Raman spectroscopy for rapid diagnosis of lymphomas and metastatic lesions found in lymph nodes

Submitted by Leanne Marie Fullwood to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Physics in September 2017.

This thesis is available for library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has been previously submitted and approved for the award of a degree by this or any other university.

(Signature).....

Abstract

At least 50% of people will develop cancer at some point during their lifetime and half these will end in fatality. Improving patients' prognosis relies on early and accurate diagnosis and treatment. Current diagnostic methods are based on histopathological analysis and are time-consuming, expensive and require biopsy. Raman spectroscopy can measure subtle biochemical changes and provides a rapid, non-destructive and objective technique that can be used *in vivo* for identifying pathological changes in tissue samples.

This study investigates both a standard Raman spectrometer system and also a Raman needle probe for their use as diagnostic techniques and clinical tools. Oesophageal, femoral and head and neck lymph nodes were analysed in this study. Metastatic lymph nodes from the three areas could be identified from the non-cancer lymph nodes with a sensitivity of 71% and specificity of 89%. Lymphoma was identified from non-cancer lymph nodes with a sensitivity of 64% and specificity of 86%. It was observed that oesophageal nodes often contained carbon particles, clinically diagnosed as anthracosis. These nodes were much harder to study than the femoral or head and neck, due to strong Raman signal detected from the carbon particles. Lymph nodes are embedded in adipose tissue and as a consequence, very strong lipid peaks were frequently observed in spectra. Spectral differences were exhibited in the measurements of the lymph nodes from the three different anatomical regions.

A comparison of the point measurements and mapped data showed no difference in classification. Therefore, indicating that just a few measurements can be sufficient enough sampling to represent a specimen, and demonstrates the practicability of Raman use *in vivo* for rapid analysis. The Raman needle probe feasibility study showed its potential for *in vivo* use for real-time diagnosis and as a surgical tool to support biopsy. A sensitivity and specificity of 80% and 79% for the identification of non-cancer head and neck lymph nodes from non-cancer provides similar accuracies to the standard Raman approach, therefore supports its viability for use as a diagnostic tool.

Acknowledgements

Firstly I would like to express my gratitude to my incredible supervisors Prof Nick Stone and Dr Catherine Kendall. Without whom I might not have had the opportunity of being part of such an interesting and evolving research area. I am so grateful for their extensive knowledge and for all of the time they have put into helping me develop as a researcher and for always getting me to ask more questions. I am extremely thankful to the LINC charity for providing the funding for this project and always allowing me to get involved in fundraising activities.

I would also like to thank Dr Gavin Lloyd for all the patience and energy he spent in explaining what he probably thought was the simplest of questions, even after every time I asked. Without him my knowledge would have been so much more limited, especially regarding chemometrics where he poured so much time into improving my data analysis.

I am also so thankful to all the surgeons at Gloucester Royal Hospital who provided the provisions and samples for me to do my research, Dr Jonathan Frost, Dr Oli Old, Mr Charlie Hall, Mr Mark Singh, Mr Hugh Wheatley, Mr Daryl Godden and Mr Mike Thomas. Also to the nurses for tirelessly informing me of the surgeries and helping me collect the samples. Thank you so much to the pathologists Dr Suzanne Hopcroft and Dr Chandima de Cates for giving me your time and knowledge for the diagnosis of the samples. A massive thank you to Prof Hugh Barr for all the humour and expertise he provides in the biophotonics unit and for all the times he's chased up people on my behalf. I will miss your kindness, generosity and intellect.

I am so grateful to have been part of an amazing research group and to those that helped me along the way; including Dr Francesco Palombo, Dr Ingeborg Iping-Petterson, Dr Krupakar Nallala, Dr Benjamin Gardner, Dr Martin Isabelle and especially Dr Alex Dudgeon for checking my thesis on multiple occasions.

I am so humbled to have been part of such an amazing group of researchers both in Gloucester and Exeter and thankful to have worked and shared offices with the most loveliest and smartest of people over these past 4 years. Thanks so much to Louise for the cooking lessons, Fay for the gym sessions and to Jen, Kelly, Martha, Dave, Toby, Michael, Beth, Skye, Rich and of course Ryan for the trips, chats and adventures. I cannot thank everyone, even though they all deserve it, but I hope they know how much I appreciate them all and am going to miss them in my day-to-day life.

Dr Sahand Zanjani pour, I am grateful not only for your excellent coding ability, but for being there for me over these past few years… and especially these last ten months.

Finally I would like to express my appreciation to my family both in Exeter and Cheltenham for supporting and loving me over the past four years. I am so grateful to my parents, sisters and brother-in-law for helping me grow in order to achieve what I've needed. Also to my little nieces for all the times they've let me get lost in their world when I've needed to. I'm so appreciative to my Auntie Dawn, Uncle Brent and cousins for allowing me to be part of their family whilst I've been studying up in Gloucester.

Contents

List of Figures

List of Tables

Abbreviations List

A	Adenine
a.u.	Arbitrary units
A/D	Analogue to digital converter
AC	Adenocarcinoma
B-cells	Bursa derived cells
C	Cytosine
CCD	Charged coupled device
CNS	Central nervous system
CT	Computed tomography
DA	Discriminant analysis
DFA	Discriminant function analysis
DNA	Deoxyribonucleic acid
ECM	Extracellular matrix
FNA	Fine needle aspirates
FNAC	Fine needle aspiration cytology
FPA	Focal plane array
FTIR	Fourier transform infrared spectroscopy
G	Guanine
H&E	Haematoxylin and eosin
HCA	Hierarchical cluster analysis
HL	Hodgkin's lymphoma
HRS	Hodgkin and Reed/Sternberg
IR	Infrared
ITC	Individual tumour cells
LDA	Linear discriminant analysis
LN	Lymph node
m	Atomic mass
met	Metastasis
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid
NHL	Non-Hodgkin's lymphoma
NK	Natural killer
PC	Principal component
PC	Papillary carcinoma
PCA	Principal component analysis
PCR	Polymerase chain reaction
PET	Positron emission tomography
PLSDA	Partial least squares discriminant analysis
PTFE	Polytetrafluoroethylene
QE	Quantum efficiency

RBF	Radial based function
REAL	Revised European American lymphoma
REC	Research ethics committee
RNA	Ribonucleic acid
ROC	Receiver operating characteristics
ROI	Region of interest
RS	Raman spectroscopy
RTPCR	Real time polymerase chain reaction
S/N	Signal-to-noise ratio
SCC	Squamous cell carcinoma
SORS	Spatially offset Raman spectroscopy
SVM	Support vector machine
T	Thymine
TAG	Triacylglycerol
T-cells	Thymus derived cells
UDR	Underlying ratio
UPC	Unknown primary cancer
VIP	Variable importance projection
WHO	World health organisation

Introduction

In the UK, people who are born after 1960 have a 50% chance of developing cancer at some point in their lifetime.[1] Approximately half will die as a result of the disease within 5 years of diagnosis. Cancer is responsible for 29% of deaths in the UK and it is estimated that at least one person dies from the disease every four minutes.[2]

Early and rapid diagnosis of disease is essential for increasing survival rates and improving prognosis. It also provides economic benefits through the relief of hospital resources and the reduction in healthcare costs.[3] Currently the method of choice for cancer diagnosis is histopathological analysis of stained tissue sections. This method requires tissue samples to be acquired from patients, typically through invasive biopsy. The unnecessary removal of tissue can result from sampling errors. This is a costly procedure and the time taken for diagnoses to be made can also be lengthy,[4] thus delaying treatment. Diagnosis is made based on observations of the tissue morphology and staining; thus, this method can be subjective and may result in incorrect diagnosis and treatment.[5] Diagnosis of a disease often only occurs as a result of the presenting of symptoms. However, by the time symptoms present, tumours tend to be of a significant size and curing the cancer completely is more difficult.[6]

There is a need for a diagnostic technique that can rapidly, non-invasively and non-subjectively identify the pathology of tissues, whilst being accurate and cost-effective.

Vibrational spectroscopic techniques are rapid, non-destructive and cost-effective methods of diagnosis. They are also minimally invasive when used *in vivo* and are non-subjective.[7] When a cell becomes diseased biochemical changes occur, which are reflected by alterations to its spectral fingerprint when compared to normal tissue spectra.[8] Vibrational techniques comprise Raman and infrared spectroscopy and function based on the interaction of optical light and matter to generate molecule specific spectra. Raman spectra are not significantly affected by water bands; therefore the technique offers the advantage of *in vivo*

use through the employment of optical fibres,[9] enabling immediate diagnosis and treatment with minimal invasion.

Although previous studies have been carried out on lymph node cancers, limited research has been conducted on primary lymphomas. There is a need for a larger study to analyse many of the different cancers that can occur in lymph nodes, including both primary and secondary. A study involving a larger range of pathologies will allow spectral comparisons to take place over many different cancers, enabling the evaluation of Raman spectroscopy as a diagnostic tool for the classification of lymph node cancers. The ability of Raman spectroscopy to correctly stage lymphomas will also be investigated, as the stage of a cancer often determines patient prognosis. The *in vivo* use of Raman spectroscopy is the ultimate aim of this technique's use diagnostically. Thus, a needle probe will be used to collect spectra *ex vivo* and analysed to evaluate its diagnostic potential for *in vivo* lymph node cancer classification.

In order to achieve the aims of the study, tissue will be obtained and analysed using vibrational spectroscopy. The spectra produced will then be assessed in order to evaluate the differences between the classification and stage assigned to the tissue section by standard histopathological assessment. Therefore, the ability of Raman spectroscopy to form pathological classifications will be assessed. The tissue that will be used in this study will be lymph node and the pathologies will consist of non-cancerous, lymphoma and metastatic tissue.

Chapter 1: Background and Literature Review

1.1 Cancer

1.1.1 Cancer Epidemiology

Currently in the UK, approximately 33%, of people will be diagnosed with cancer at some point during their lifetimes and around half of these will die as a result of the disease within 5 years of diagnosis. This occurrence is expected to rise to 50% of people. Cancer is the cause of death in one in every four fatalities in the UK and in 2010 it was estimated that cancer took a person's life every four minutes.[10] More than 60% of cancers are typically diagnosed in people who are over 65 years old. In 2006, 250,000 overall cancer diagnoses and 120,000 cancer-caused fatalities were estimated in the UK.[11]

Cancer represents a group of over 100 diseases which are characterised by their uncontrolled growth, invasion and cell dissemination; and classified according to their tissue of origin. Carcinomas are cancers that occur in epithelial cells and account for roughly 85% of overall cancers. Adenocarcinoma is the collective name for cancers that occur in glandular epithelial tissue, such as that found in the breast. Mesoderm cells give rise to bone and muscle tissue; cancers that occur in mesenchymal tissue are referred to as sarcomas.[12] Cancers that occur in blood forming cells and those from the immune system are either leukaemia or lymphomas and account for 7% of all cancers.[13] Cancers that occur in glial cells of the brain are collectively called gliomas.[14]

It is estimated that up to 90% of cancers are caused by environmental or lifestyle factors and are often the initial trigger for tumour development; genetics also is known to contribute to the growth of neoplasms.[11]

1.1.2 Cancer Pathogenesis

A constant equilibrium between cell proliferation, growth arrest and differentiation, and apoptosis is successfully controlled through homeostasis in

normal tissue. Alterations to the homeostatic mechanisms that control this equilibrium can result in an increase in tissue mass and thus tumour formation.[15] Cell proliferation is a mechanism needed for the normal growth and development of organisms, where cells are exponentially produced through DNA synthesis followed by mitosis. A typical, healthy human cell can divide once or twice a day, replication is controlled to prevent over-proliferation. Differentiation is tightly regulated in normal tissue; directly related to a cells ability to proliferate, it refers to the specialisation of cells. Apoptosis is a form of cell death and essential for homeostatic maintenance in healthy tissue. In neoplastic tissue the following cellular functions are typically regulated incorrectly: cell proliferation is not controlled effectively, cell differentiation is distorted, chromosomal and genetic organisation is weak and apoptosis is inefficiently monitored.[16] Telomeres are repeated sections of DNA that protect chromosomes. Alteration or shortening of telomeres can result in a loss of this protection and can result in cell apoptosis and chromosome fusion. Telomerase can be expressed by cells to reverse telomere shortening through the addition of nucleotides. It has been shown that over expression of telomerase results in greater cell mobilisation and a greater life span, leading to a higher ability of tumour initiation.[17]

Hanahan and Weinberg[18] propose six hallmarks of cancer cells: sustained proliferative signalling, evasion to growth suppressors, activation of invasion and metastasis, enabled replicative immortality, induced angiogenesis, and resistance to cell death. They state, that as a normal cell progresses into a neoplastic state these hallmark capabilities are acquired successively to enable them to become tumorigenic and eventually malignant. They explain that normal cells control cell numbers through maintenance of the generation and release of growth-promoting signals; cancer cells sustain their proliferation by deregulating these signals. Tumour-suppressor genes negatively regulate cell proliferation; cancer cells have the ability to evade their suppression. Contact inhibition aids proliferation in normal cells, this mechanism is not present in many cancer cells allowing invasion of tissue. Angiogenesis is the formation of new vessels from existing ones, needed in cancerous tissue to provide oxygen and nutrients and remove waste to sustain neoplastic growth. Apoptosis is a method of cell death whereby the dying cells contract and are engulfed by neighbouring cells; cancer cells are able to evade this mechanism.

1.1.3 DNA Mutation

Deoxyribonucleic acid (DNA), a double stranded helix structure, is formed from a sugar-phosphate backbone attached to nitrogenous bases responsible for the specific genetic code. The backbone is composed of sugar units of deoxyribose connected together by a phosphate group. A 3'-hydroxyl group is free at one end of the chain and a terminal 5'-hydroxyl group is found at the other end with opposing polarity; this results in the two DNA strands of the double helix running in opposite directions. Nitrogenous bases include the purines: guanine (G) and adenine (A), and the pyrimidines: thymine (T) and cytosine (C), which complement each other G:C and A:T. Amino acids are coded for by a codon which consists of three bases. RNA polymerase, the enzyme that copies DNA into messenger mRNA after the strands of the double helix are separated, reads in a 5'-3' direction. Base mutations occur from misreading of the strands during synthesis and occur in three ways: single base alteration, insertion of a base, or deletion of a base. DNA mutations can result in repair and return to the normal sequence, cell death, or the continued misreading of the mutated DNA strand in subsequent DNA syntheses. Cancer occurs when the damaged DNA sequence codes for a gene, the coding sequence of a protein, and thus alters cellular function and imbalances homeostatic maintenance which leads to tumour formation.[19]

1.1.4 Carcinogenesis

Carcinogenesis is a multistage process whereby normal cells progress into cancerous cells, and is typically divided into three stages: initiation, promotion and progression (Figure 1.1).[20]

In the initiation step, exposure to a chemical carcinogen, radiation or a virus results in a mutagenic effect that causes damage to the DNA of a cell. This is an irreversible process, occurs with no change in outward appearance and dramatically increases a cells suceptibility to develop cancer. Chemicals capable of causing initiation of a cell are highly reactive electrophiles, where the electron-deficient atom can react with the electron-rich nucleophilic sites within cells, a

primary target being DNA or RNA. Promotion is a reversible process where a promoting-agent does not directly affect DNA, but causes proliferation of the cell and angiogenesis. Promoters are non-mutagenic and do not typically cause tumours themselves, but stimulates the initiated cell to undergo cell division and produce more DNA-mutated cells. Thus, promotion aids tumour progression through proliferation, and results in the formation of a benign tumour. A long interval may occur between initiation and promotion due to the requirement of a prolonged promoting-agent exposure. The progression stage involves irriversible processes such as de-differentiation and an increase in autonomous growth and mitotic activity. The tumour becomes malignant, can invade local tissue and has the ability to metastatisise to secondary locations around the body.[11,15,19,21]

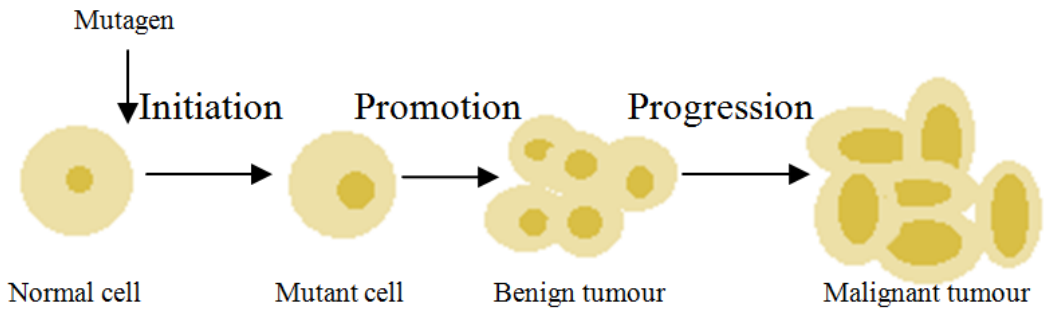

Figure 1.1- The basic process of carcinogenesis. Adapted from [22].

A metastasis occurs when tumour cells detach themselves from the malignant site and travel to other secondary locations where they form deposits and proliferate.[23] All cancers have the ability to metastasise and can do so by various routes after cell detachment; invading local tissue, seeding into surfaces and body cavities, haemotogenous spread or lymphatic spread.[24] Figure 1.2 shows how a primary tumour may result in metastatic deposits in the lungs. The cell of origin of a tumour affects metastatic proclivity; the cells of connective tissue cancer typically migrate independently whereas carcinomas typically migrate collectively. Cancer patients tend to die as a result of the metastases rather than the primary cancer. It is estimated that every day, for every gram of tumour tissue, four million cancer cells are released into the circulatory system.[25] Within twenty-four hours of the cancer cells entering the blood stream only 0.1% remains viable,

and of those, approximately only 0.01% will eventually produce metastases. Subsequently, only a small proportion of primary malignancies result in the formation of metastases at secondary locations.[26] Tumour cells are genetically less stable than normal cells [27] and are highly heterogeneous. Comparative genomic hybridisation has found that there is a high level of similarity between brain metastases and their primary site.[28] However, it has also been reported that cells which possess a higher metastatic potential, measured using in vitro assays to calculate the proliferation, also have a higher mutation capability than those without such a high metastatic potential. Thus cells at secondary locations may genetically differ from their primary site cells.[26]

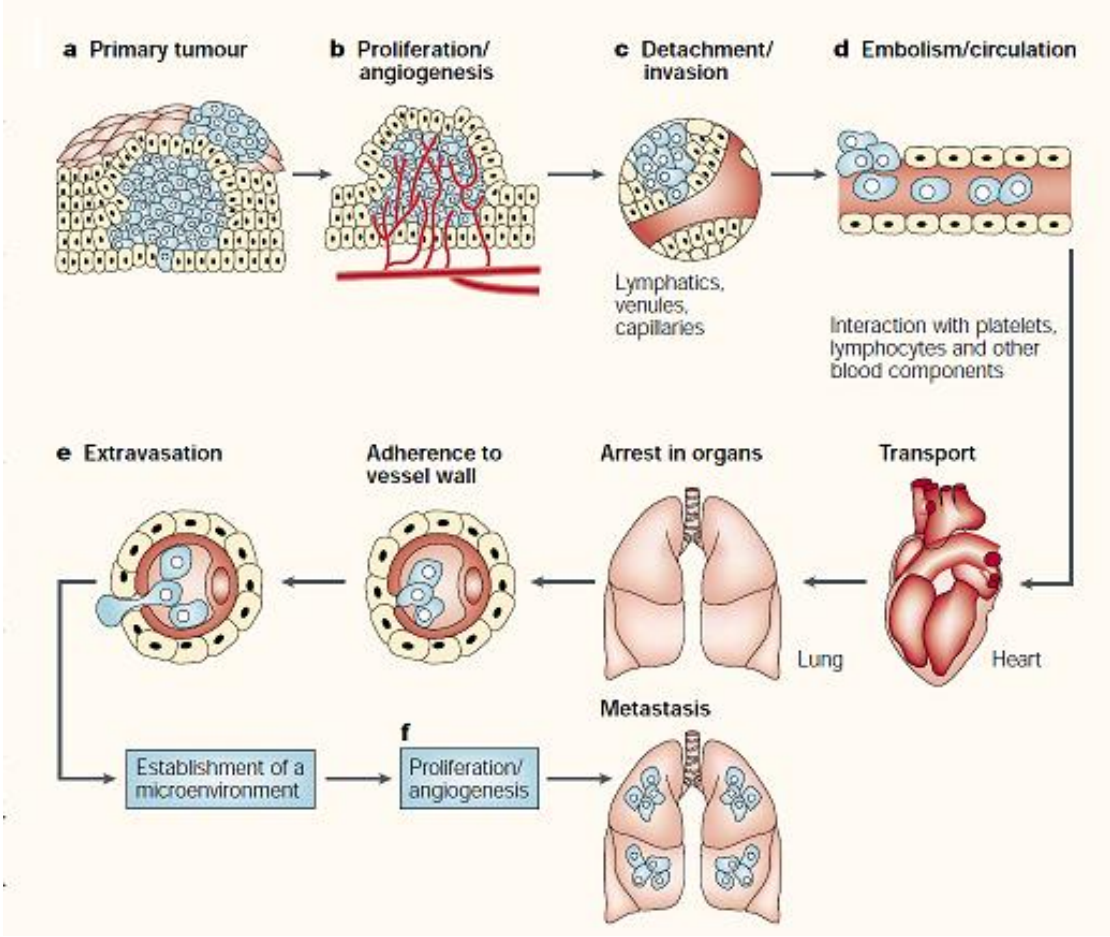

Figure 1.2- Schematic showing the process of metastasis. Replicated from reference.[26]

1.1.5 Cancer Cell Morphology

Increased cell growth and proliferation that occurs in cancerous cells can be recognised by localised and progressive tissue expansion. This is caused by an increase in the number of cells as a result of proliferation. Cancerous epithelial cells exhibit pleomorphism that can be recognised under a light microscope. Morphological differences of cancerous cells compared to non-cancer cells include: altered nuclear size and shape, the distribution of chromatin and staining affinity. Invasion and metastasis of cells are also indications of cancer. Invasion is the random and uncontrolled penetration of the cancerous cells into surrounding tissue. Metastasis is demonstrated by the dissemination of cancerous cells from the site of origin to secondary sites in the body.[29] Morphological differences also reported include: variable cell shape and size, no basement membrane, lack of tissue organisation, altered epithelium/stroma ratio, invasion of blood vessels and lymphatic vessels, increased and abnormal mitotic activity and a loss of differentiation.[19]

The conformations and quantities of: nucleic acids, proteins, lipids and carbohydrates are also altered when a normal cell becomes neoplastic. These morphological and biochemical changes vary between the type and location of the cancer.[30]

Benign and malignant tumours need to be correctly identified to ensure the correct treatment. Benign tumours, though not typically fatal, can cause health problems depending on their anatomical location. Malignant tumours progress and usually result in mortalities if the tumour is unnoticed and is not treated or if the treatment is unsuccessful. Microscopic observation of invasion into surrounding tissue and tumour metastasis to secondary locations primarily identifies malignant cancers from benign tumours.[31]

1.2 Lymph Node Cancers

1.2.1 Lymphatic Anatomy and Functionality

The lymphatic system consists of a network of vessels throughout the body which flow with a fluid called lymph (Figure 1.3). Lymph fluid is derived from the interstitial fluid that surrounds tissues and originates in the interstitial spaces between tissues.[32] There is both primary and secondary lymphoid tissue. Primary lymphoid tissue consists of the bone marrow, where B cells and T cells arise and B cells differentiate into mature B-cells, and the thymus where T cells differentiate into mature T-cells. These mature B and T cells express antigen receptors and migrate into the secondary lymphoid tissues. The secondary lymphoid tissues include the lymph nodes, the spleen, mucosa associated lymphoid tissue and the blood.[33]

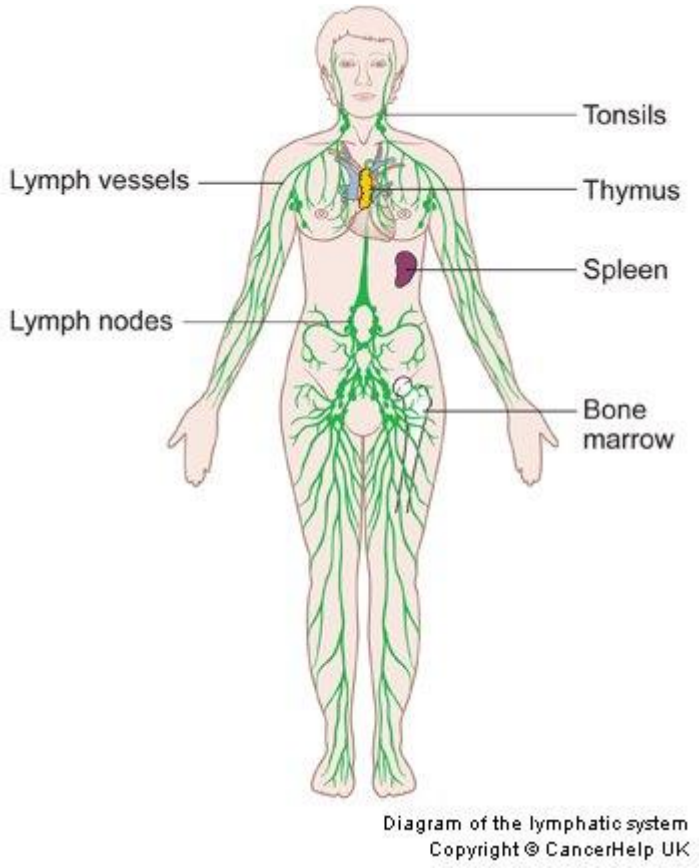

Figure 1.3- The lymphatic system and associated organs in the Human body. Reproduced from reference.[34]

The lymphatic system is involved in the body's immunity against foreign microbes. The tissues of the lymphatic system, such as the spleen and lymph nodes, contain macrophages, a phagocytic cell that engulfs foreign bodies, to enable pathogen destruction. Microbes from the blood are trapped in the spleen and those present in interstitial fluid are taken into the lymphatic vessels and travel to lymph nodes for destruction by macrophages. Fluid from the circulatory system escapes the vessels as a result of blood pressure. Through osmotic pressure, approximately 85% of fluid re-enters the circulatory system. However, 4 L of fluid is lost every day from the capillaries into the interstitial fluid and does not re-enter the circulatory system. Through diffusion this fluid enters the vessels of the lymphatic system and is transported by rhythmic movement of the vessel walls and contractions of the skeletal muscles. Similarly to veins, valves are present on the walls of the lymphatic vessels to prevent backflow.[35]

As well as extravasated fluid, the lymphatic vessels also collect macrophages and other cellular material related to the immune system. The lymphatic network begins as capillaries that drain into collecting vessels which return lymph fluid to the systemic blood circulatory system through the thoracic duct at the base of the neck. The lymphatic system is lined with endothelial cells and is designed for cell entry and transportation through a variety of characteristics: Lymphatic vessels contain no basement membrane and few intracellular junctions, aiding cell entry; cells are easily washed into the lymphatic vessels with interstitial fluid and proteins; the flow velocity in the lymphatic network is slow and the chemistry of lymph is similar to that of interstitial fluid, thus generates an environment that facilitates cell survival. Therefore, the lymphatic system aids the body's immunity by taking lymphocytes and antigen-presenting cells to the lymph nodes. However, it can also act as a transportation system for the dissemination of tumour cells and thus aids metastasis. The presence and over-expression of specific lymphangiogenic growth factors may contribute to the migration of tumour cells by providing a suitable environment. The over-expression of the family of vascular endothelial growth factor proteins facilitates lymphatic and vascular proliferation.[36]

Lymph nodes are located at positions along the vessels and filter lymph fluid; they can be felt in the neck, armpit and groin region. Lymph nodes consist

of a ‘honeycomb’ of connective tissue containing white blood cells to aid the body’s immunity. When the body is fighting infection white blood cells divide rapidly and thus the nodes can be swollen. They can be assessed in cancer patients to enable the monitoring of metastasis.[35]

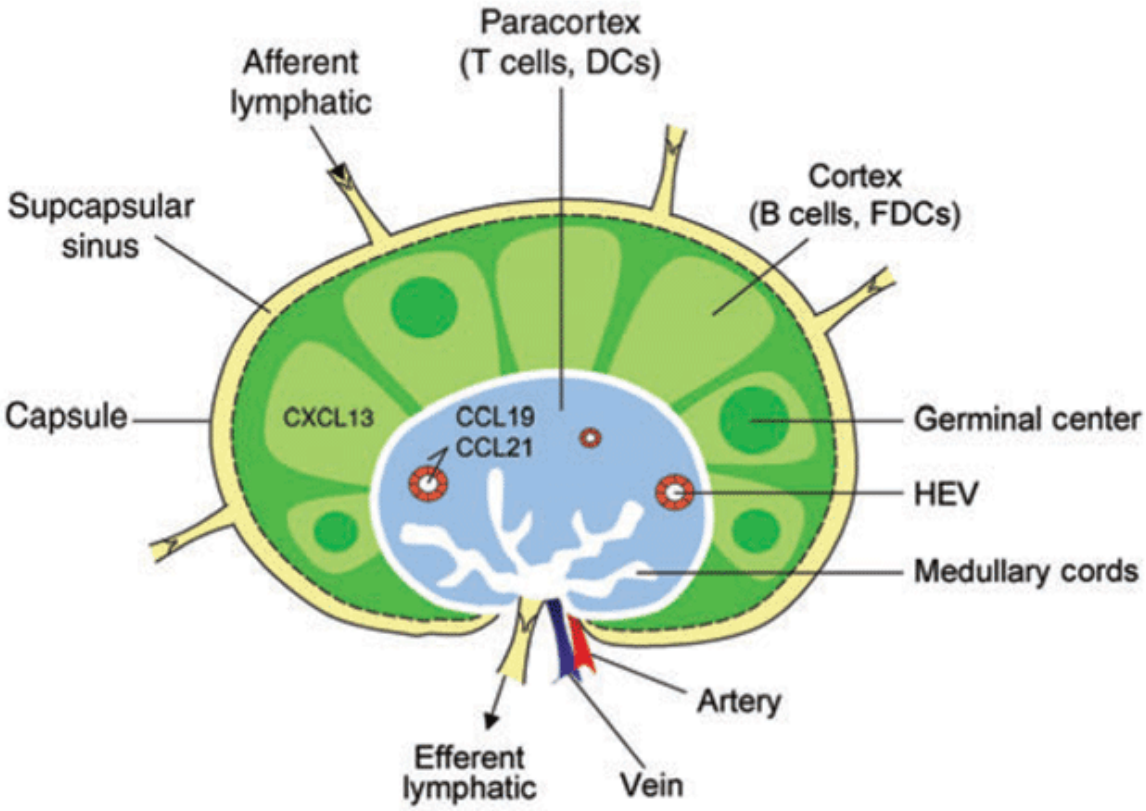

Figure 1.4- A diagram of the structure of a lymph node. It can be divided into three primary areas: cortex, paracortex and the medulla. Reproduced from reference [37].

Lymph nodes are secondary organs in the lymphatic system. Their structure consists of a cortex with lymph follicles (where B-cells are located), a paracortex (where T-cells are located), and a medulla (Figure 1.4). Each compartment contains specific subsets of T lymphocytes and B lymphocytes, macrophages, dendritic cells and stroma cells.[38] Lymphocytes are a type of white blood cell produced in the bone marrow. T lymphocytes are responsible for cell-mediated cytotoxic reactions and for mediating inflammatory reactions. T cells also produce the cytokines that regulate the immune response and provide helper activity for B cells. B lymphocytes are the precursors of immunoglobulin-secreting plasma cells and they concentrate and present antigens to T cells. Resting

lymphocytes are typically small round cells with a dense nucleus and little cytoplasm in appearance.[39] Lymph nodes are typically 8-12 mm in diameter but can vary greatly depending on the individual and anatomical location.[40]

Lymph nodes are found where vessels converge; they are discrete masses of fibrovascular tissue enclosed by the lymphatic vessel. Approximately 450 lymph nodes are found in the human body. Lymph nodes are formed during development when a mesenchymal bud invaginates into a lymphatic sac and compresses the lumen of the sac against the opposite wall. The sac becomes covered by the endothelium lining as it expands. The sac wall becomes the capsule and the point at which invagination occurs becomes the hilum. A number of lobules are found within lymph nodes. The mesenchymal tissue differentiates into these lobules and the sac lumen develops into the system of sinuses that surround the lobules; the capsule encloses these structures. The lobules collect inflammatory mediators and dendritic cells (antigen-presenting cells) from lymph which flows through the system of sinuses, exposing its content to lymphocytes.[32] After monocytes are released from the bone marrow into circulation they migrate into various tissues where they differentiate into dendritic cells or macrophages. Macrophages are found in almost all tissue of the body and respond to external stimuli as a result of infection to aid the body's immunity.[41] Receptors on their surface detect the signals that are not usually present in normal cells. Macrophages in lymph nodes remove viruses from lymph and initiate antiviral humoral immune responses. Therefore, macrophages both fight infection and remove dead cells and toxic materials. They also restore tissue homeostasis through reducing the inflammation following infections.[42] The lymphoid lobule is the basic functional unit of the lymph node and their number varies between nodes. Its anatomical structure provides indications regarding the functionality and pathology of the lymph node. The primary immune response occurs in the lymph nodes. In the reticular meshwork of the stromal structure, antigens are displayed by antigen-presenting cells to lymphocytes, which then undergo clonal expansion to produce new lymphocytes and plasma cells to secrete and release antibodies into the lymph. When foreign antibodies invade the body, a local immune response is activated at the site of infection. The lymphatic vessels

collect the antigenic material, dendrite cells and inflammatory mediators present at the site of infection and transport them in the lymph fluid.[32]

1.2.2 Cancers of the Lymph Nodes

Cancers of the lymph node can be either primary neoplasms, which have originated from the cells of the lymph node, or secondary neoplasms, where cells from tumours in other organs metastasise to the lymph nodes and form deposits. Primary cancers of the lymph node are primarily classified as either Hodgkin's or non-Hodgkin's lymphoma, and constitute the most common type of blood cancer.[43]

In 1832 Thomas Hodgkin first recognised human lymphomas, leading to the classification of Hodgkin lymphoma and their distinction from other types of lymphomas. Classification is based on both morphology and molecular constituency of the tissue.[44] Lymphomas are types of blood cancer and are primary lymph node cancers, which originate from lymphocyte cells found in the nodes. Lymphomas are divided into Hodgkin's lymphoma (HL) and non-Hodgkin's lymphoma (NHL). According to the Cancer Research incident and mortality UK reports for 2010,[10] NHL was the 5th most common cancer in men and the 7th most common cancer in women. Of those diagnosed, 37% of males and 36% females died as a result of the disease. In males, NHL accounted for approximately 8% of the total cancers diagnosed in 25-49 year olds, 4% in 50-74 year olds and 4% in 75 year olds and over. A 12% increase of NHL diagnoses has been observed from 1999-2010; however, the death rate has decreased by 15% in males and 21% in females. Hodgkin's lymphoma is much more infrequent with only 17% of those diagnosed resulting in fatalities. In males, an 11% increase in HL diagnosis has been observed from 1999–2010.

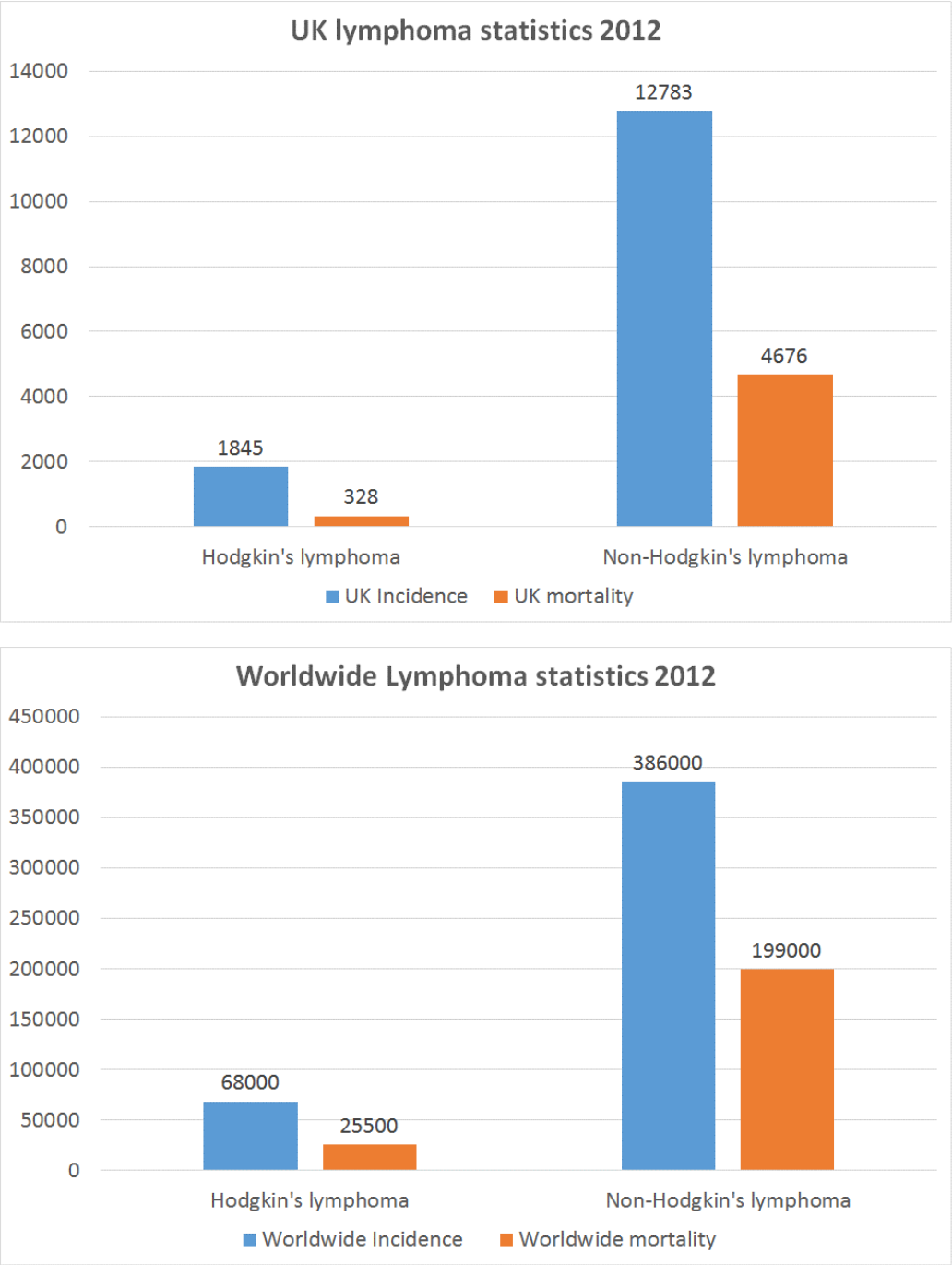

Figure 1.5- A bar graph to show the number of diagnoses and fatalities of non-Hodgkin's and Hodgkin's lymphoma in the UK (top) and worldwide (bottom).[45]

Hodgkin's lymphoma will almost always present as a nodal disease with a propensity for occurrence in head and mediastinal nodes. However, in up to 40% of non-Hodgkin's lymphomas, the disease will evidence itself in an extranodal location.[46] Extranodal lymphomas are those that arise from tissue other than the lymph nodes. The most frequent site for extranodal non-Hodgkin's lymphoma is the gastrointestinal tract, being the location for 30-40% of this type of lymphoma.[47] Extranodal lymphomas associated with the skin or central nervous

system (CNS) are also common. They often originate from mucosa-associated lymphoid tissue and are identified if there is little or no involvement of the lymph nodes and there is a clinically dominant extranodal component. The incidence of extranodal lymphomas has increased in years, especially those of the CNS, thought to be attributed to the increase in viral infections, such as acquired immunodeficiency syndrome (AIDS), and the use of immunosuppressant treatments. Patients with these lymphomas are less likely to present with B symptoms (Table 1.1) than patients with lymphomas arising in lymphatic nodal regions.[48]

Table 1.1- Revised European American Lymphoma (REAL)/World Health Organisation (WHO) classification of lymphoid neoplasms.[49]

Non-Hodgkin's lymphoma
B-Cell neoplasms
Precursor B-cell neoplasm - Precursor B-lymphoblastic leukaemia/lymphoma (precursor B-cell acute lymphoblastic leukaemia) **Mature (peripheral) B-cell neoplasms** - B-cell chronic lymphocytic leukaemia/small lymphocytic lymphoma - B-cell prolymphocytic leukaemia - Lymphoplasmacytic lymphoma - Splenic marginal zone B-cell lymphoma (1/2 villous lymphocytes) - Hairy cell leukaemia - Plasma cell myeloma/plasmacytoma - Extranodal marginal zone B-cell lymphoma of MALT type - Nodal marginal zone B-cell lymphoma (1/2 monocytoid B cells) - Follicular lymphoma - Mantle-cell lymphoma - Diffuse large B-cell lymphoma - Mediastinal large B-cell lymphoma - Primary effusion lymphoma - Burkitt's lymphoma/Burkitt cell leukaemia
T-cell and NK-cell neoplasms
- Precursor T-cell neoplasm - Precursor T-lymphoblastic lymphoma/leukaemia (precursor T-cell acute lymphoblastic leukaemia)

- Mature (peripheral) T-cell neoplasms - T-cell prolymphocytic leukaemia - T-cell granular lymphocytic leukaemia - Aggressive NK-cell leukaemia - Adult T-cell lymphoma/leukaemia (HTLV11) - Extranodal NK/T-cell lymphoma, nasal type - Enteropathy-type T-cell lymphoma - Hepatosplenic gamma-delta T-cell lymphoma - Subcutaneous panniculitis-like T-cell lymphoma - Mycosis fungoides/Sezary syndrome - Anaplastic large-cell lymphoma, T/null cell, primary cutaneous type - Peripheral T-cell lymphoma, not otherwise characterized - Angioimmunoblastic T-cell lymphoma - Anaplastic large-cell lymphoma, T/null cell, primary systemic type
Hodgkin's lymphoma (Hodgkin's disease)
Nodular lymphocyte-predominant Hodgkin's lymphoma **Classical Hodgkin's lymphoma** - Nodular sclerosis Hodgkin's lymphoma (grades 1 and 2) - Lymphocyte-rich classical Hodgkin's lymphoma - Mixed cellularity Hodgkin's lymphoma - Lymphocyte depletion Hodgkin's lymphoma

Cancer classification is based on the assumption that cancers from the same anatomical location and of the same histology will have similar patterns of development and thus will have similar prognostic outcomes.[50] Currently lymphomas are classified in accordance with the Revised European American Lymphoma (REAL)/World Health Organisation (WHO) classification system. The development of the REAL and WHO classification systems involved the consensus of 19 and 50 pathologists respectively.[51] Two lymphoid differentiation pathways occur in the immune system, either involving T (thymus-derived) cells or B (bursa-derived) cells. Therefore, non-Hodgkin's lymphomas are separated into either T-cell tumours or B-cell tumours.[52] Approximately 90% of aggressive lymphoma cases are derived from B-cells.[53] Diffuse large B-cell lymphoma is the most common subtype of non-Hodgkin's lymphoma. It is clinically heterogeneous

and approximately only 40% of patients respond to treatment sufficiently enough to prolong life.[44]

There are two classifications of Hodgkin's lymphoma, nodular lymphocyte predominance HL and classical HL. Neoplastic cells can be characterised by lineage specific markers, activation markers and transcription factors. Reed-Sternberg cells are often present in the lymph nodes of individuals with Hodgkin's lymphomas and thus the identification of these cells is essential for diagnosis of the disease. However, affected lymph nodes consist of less than 1% of these neoplastic cells.[54]

Table 1.2- The stages of lymphoma disease[55] based on the Ann Arbor method of lymphoma staging.[56]

Stage	Features
I	Single lymph node region affected/one extralymphatic site affected
II	Two or more lymph nodes affected on the same side of the diaphragm/one extralymphatic site and at least one lymph node region affected on the same side of the diaphragm.
III	Lymph node regions on both sides of the diaphragm affected and can include extralymphatic site involvement also.
IV	Diffuse involvement of at least one extralymphatic site.
Suffix	
A	No symptoms
B	The patient experiences any of the following symptoms: weight loss (>10% baseline during 6 months before staging), recurrent unexplained fever (>38^{o}C), or night sweats.
X	Bulky tumour tissue (>10cm in largest diameter or mediastinal tumour which is >1/3 of maximum transverse transthoracic, diameter measured on posterior-anterior chest radiograph)

Lymphatic vessels aid cancer metastasis by undergoing changes to assist the entry of cells into lymphatic vessels. These changes involve

lymphangiogenesis and lymphatic enlargement in the lymphatics around the primary tumour. Growth factors, produced from tumour cells and cells from the immediate environment encourage lymphangiogenesis, where new lymphatic vessels are formed from the pre-existing lymphatic network, and the enlargement of the lymphatic vessels in and around the tumour site. Enlargement of the vessels can also lead to remodelling of the lymphatics by the surrounding smooth muscles. The prognosis of a patient is decreased rapidly upon the dissemination of cancer from a primary site to the lymph nodes.[57] Metastatic cancers in the lymph nodes are those that have spread there from other organs. They can be squamous cell carcinomas (SCC), from malignant primary cancers of the upper aero-digestive tract, or adenocarcinomas, from malignant primary cancers in the head, neck, breast and oesophagus.[58] When a cancerous cell spreads from one organ and settles in another it continues to divide and although the cells are mutated, it typically recreates its original tissue where it metastasised from. Thus, the secondary tumour structure takes the form of the tissue architecture of the primary cancer.

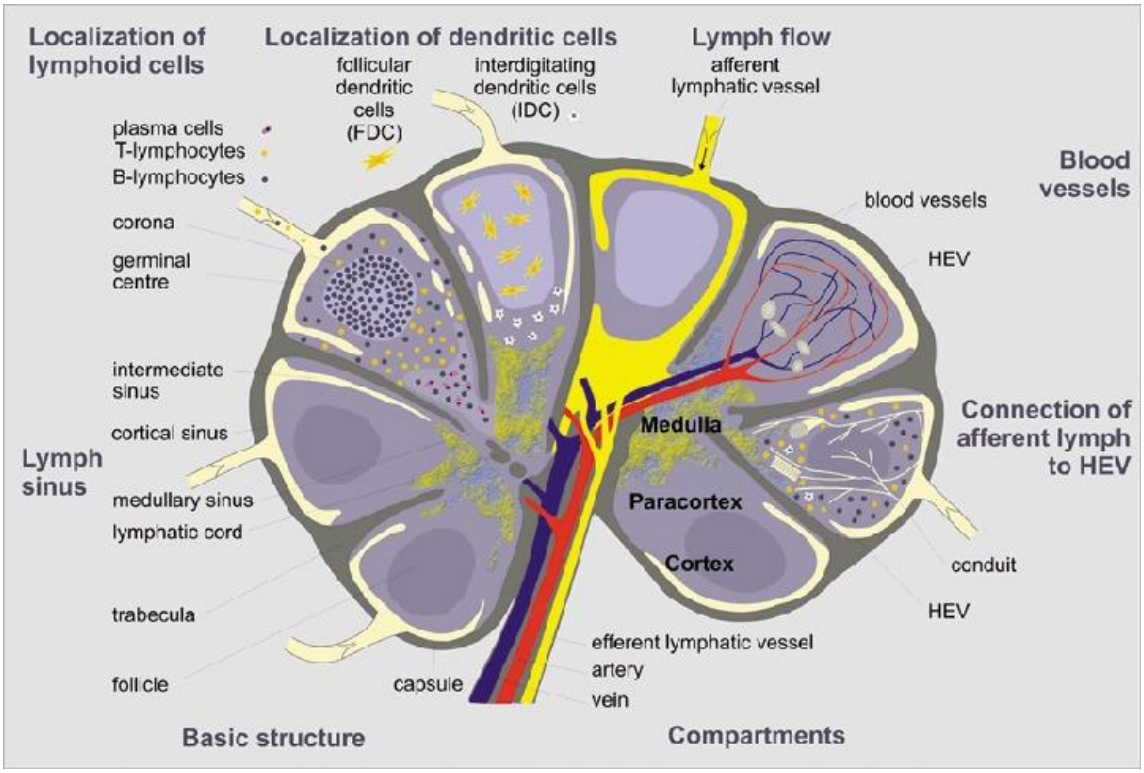

Figure 1.6- Lymph node anatomy, reproduced from reference [59].

Afferent lymphatic vessels draining connective tissue and peripheral lymph nodes empty into the subcapsular sinus. The lymph fluid then enters the medullary sinuses, via intermediate sinuses, and eventually is collected by efferent lymphatic vessels. Around the hilus the subcapsular sinus filters the lymph fluid directly into the medullary sinises. The direction of lymph flow is controlled by valves in the vessels.[60] Thus, infiltrating tumour cells first collect in the subcapsular sinus as individual cells or clusters and then infiltrates the medullary sinuses, medulla and eventually the cortex. Partial or whole lymph node replacement finally occurs. Due to the invading metastasis the lymph node undergoes changes throughout its structures, for example hyperpasia to the follicles and paracortex and sinus histiocytosis are a few. Cytologically the metastatic cells should largely be similar to the cells of the primary tumour, although not in all cases.[61]

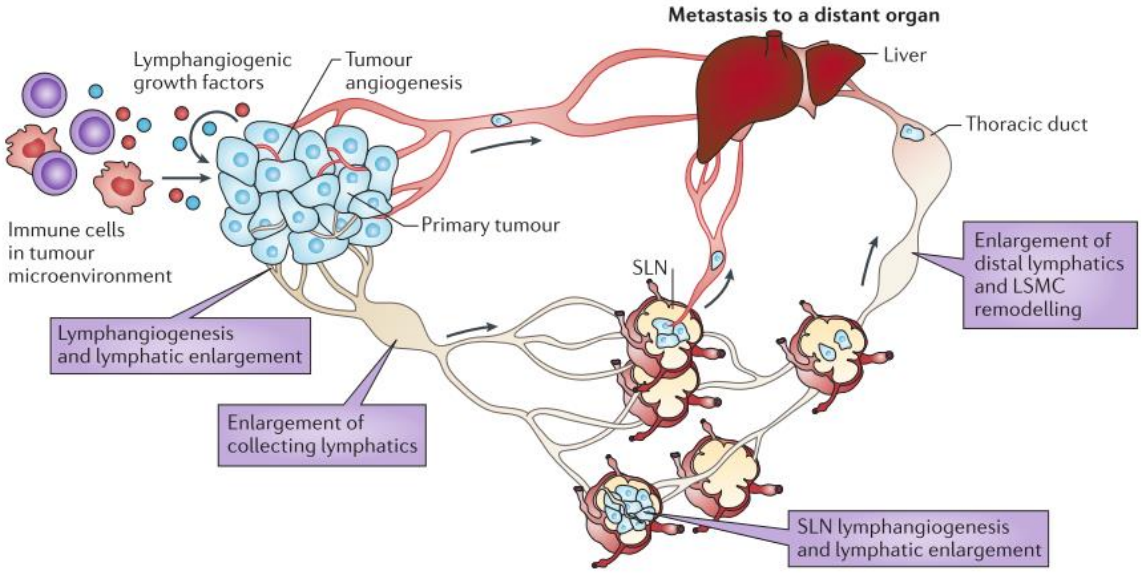

Figure 1.7- The metastatic process of tumour cells using the lymphatic system as a pathway for dissemination to lymph nodes and other secondary sites. Reproduced from reference.[57]

As primary malignancies increase in size, the frequency of regional lymph node involvement and distant metastases also increase.[62] Although the pattern of lymphatic drainage varies between anatomical locations of lymph nodes, the prognostic outlook of a patient is influenced by the site of the lymph node metastases. Prognosis of a patient with head and neck lymph node metastases is often worse if metastases occur in lymph nodes beyond the first level of

lymphatic drainage and if the lymph nodes that are involved are in lower regions of the neck, such as the supraclavicular region.[50] The presence of lymph node metastases is a useful indicator of a patient's prognosis and is often used to stage the cancer. Studies have shown that as the number of lymph node metastases increase, patients' chance of survival decrease.[63]

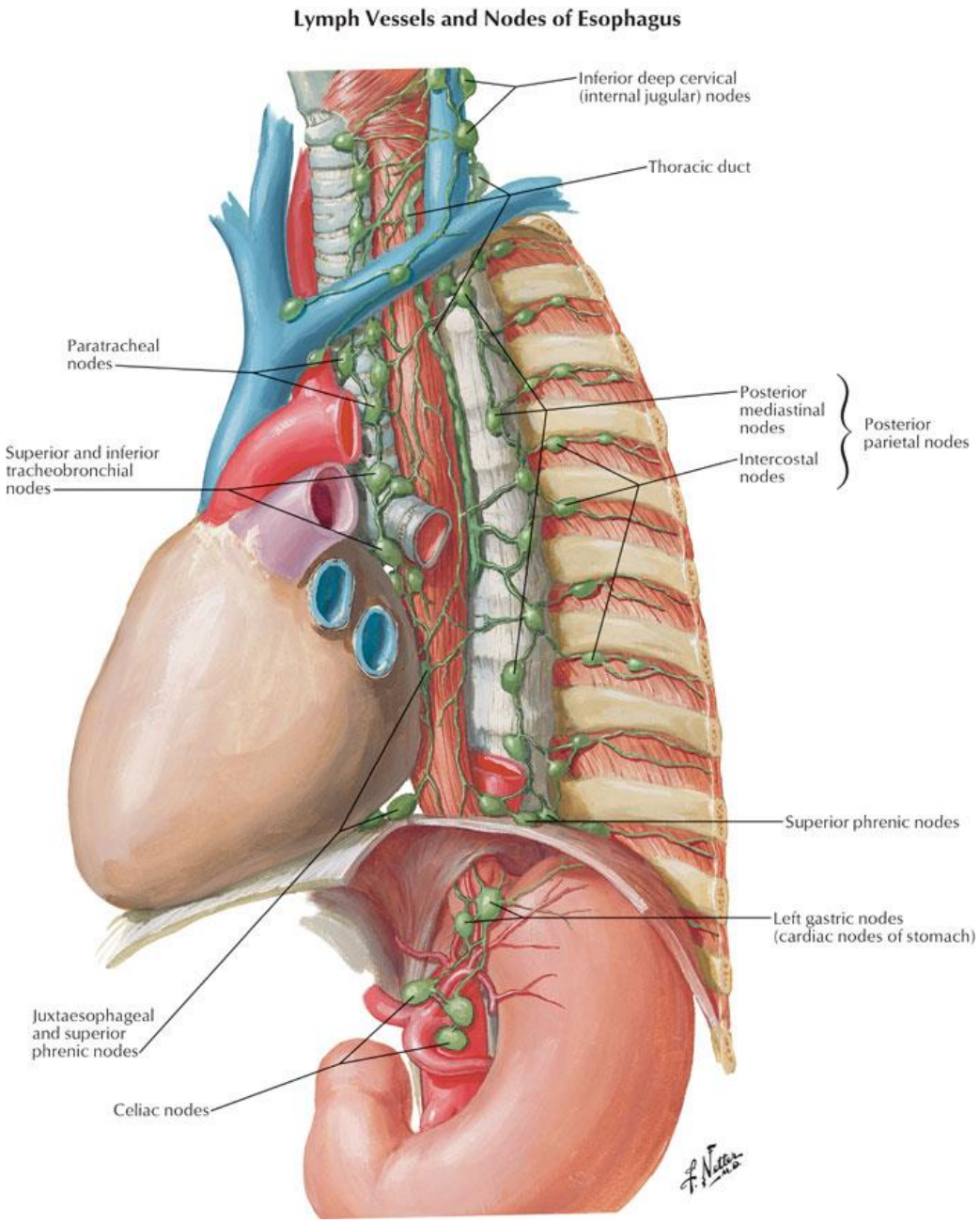

Figure 1.8- Image of oesophageal lymph nodes, reproduced from reference.[64]

Table 1.3- Head and neck lymph nodes are distinguished based on 7 anatomically distinguished sites.[50]

Lymph node level	Anatomical location
Level I	Submental
	Submandibular
Level II	Upper jugular
Level III	Mid jugular
Level IV	Lower jugular
Level V	Posterior triangle
Level VI	Prelaryngeal
	Pretracheal
	Paratracheal
Level VII	Upper mediastinal
Other groups	Sub occipital
	Retropharyngeal
	Parapharyngeal
	Buccinator
	Preauricular
	Periparotid and intraparotid

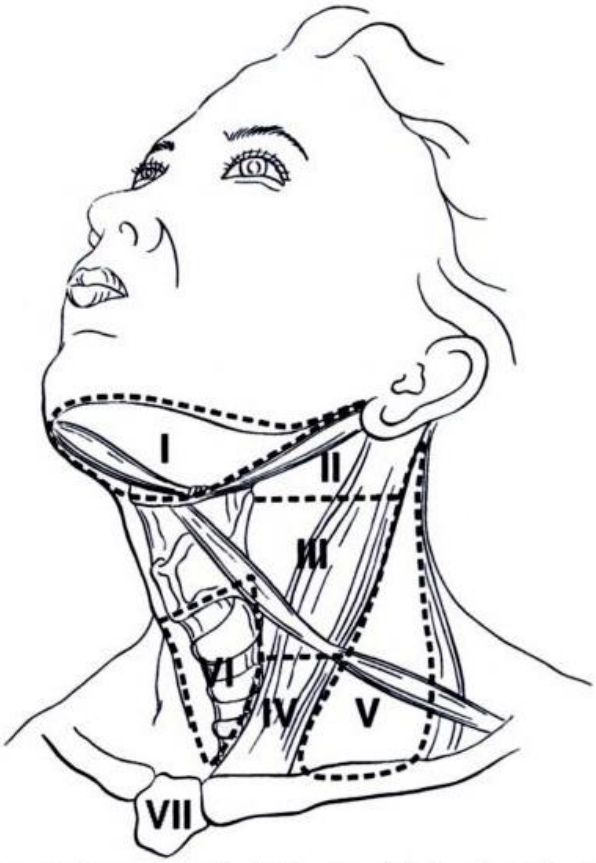

Figure 1.9- Schematic of the anatomical location associated with the classification level of head and neck lymph nodes. Replicated from reference[50].

1.2.3 Symptoms

The most frequent symptom of a lymphoma is lymphadenopathy which is the enlargement of the lymph nodes in the neck, armpit or groin. They are usually painless but can ache. Other associated symptoms, which are often rare, include: prolonged infections and a susceptibility to develop them, fatigue, night sweats, weight loss, vomiting, high body temperature, breathlessness and persistent itchy skin. These symptoms are not specific to lymphomas and infections and other conditions can produce the same symptoms.[65]

1.2.4 Pre-operative Diagnosis

Blood tests can be carried out initially to assess the level of each type of blood cell: white blood cells, red blood cells and platelets, and to examine the efficacy of the liver and kidneys. Diagnosis can also be aided by imaging

techniques such as computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET). CT is used to produce a series of two-dimensional images in order to construct a three-dimensional image. The images are produced by scanning the patient with a narrow beam of x-rays which are detected by a device directly opposite the x-ray source.[66] The variation between magnetic susceptibility of tissues is the basis for MRI. The magnetic susceptibility is a measure of a materials' ability to interact with and distort a magnetic field [67]. PET is an imaging technique that labels the small biological molecules of interest with positron-emitting radionuclides. When these positron-emitting labels undergo radioactive decay they can be detected by the PET scanner. Temporal distribution imaging can allow the formation of physiological maps. Unlike MRI and CT, PET has the potential to differentiate between benign and malignant tumour masses.[68] Imaging techniques enable non-invasive, in vivo monitoring allowing analysis of the tumour mass, its margins and its location.[69] The assessment of nodal size is limited as lymph node enlargement is not confined to cancer alone and can occur as a result of other factors, such as infection. When an upper limit in nodal size of 1 cm is used a sensitivity of ~64% and a specificity of ~62% occurs.[70] Kerr *et al.* reported that 58% of malignant and 43% of benign lymph nodes were over 15 mm in size.[71] The similar features of benign and malignant cancers also limit the diagnostic use of MRI.[70] Ultrasonography can also be used to assess tumours; however, these non-invasive pre-operative techniques cannot differentiate between benign and malignant tumours alone.[72]

These examinations are only used to preliminary assess a suspected tumour due to the risk of misdiagnosis when relied upon alone. For brain cancer, it has been reported that false positive diagnoses occur in approximately 11% of patients with brain lesions when based only on the MRI scan. More definitive, examinations must be carried out to allow more accurate diagnosis.[73] Moreover, early and accurate diagnosis is required; these methods cannot identify solid tumour masses smaller than 1 cm^3 in diameter (approximately 10^9 cancer cells).[74]

1.2.5 Post-operative Diagnosis

Currently the gold standard for cancer diagnosis is histopathological analysis of excised specimens, where the architecture of the tissue is microscopically assessed. In order for this method of analysis to be performed the tissue must be sectioned and stained, therefore requiring tissue removal. The enlarged lymph node is removed in cases of suspected primary lymphatic cancers. For cases of suspected metastasis, a sentinel lymph node biopsy is conducted. It is carried out on the basis that cancer cells that have invaded the lymphatic system from external locations, will have initially passed through sentinel nodes, as these nodes are closest to the tumour. Sampling the sentinel nodes is therefore a less invasive method of identifying metastatic spread, than sampling all regional nodes. However, in approximately 25–30% of cases the result from sentinel lymph node analysis will be positive for metastases and a subsequent operation will be required to remove affected lymph nodes. This delays treatment, uses up hospital resources and can cause psychological upset to the patient.[75] Unnecessary removal of the lymph nodes can cause unneeded health problems for the patient such as lymphoedema. Lymphoedema is caused when the lymph nodes that would otherwise filter lymph fluid have been removed, causing the fluid to build up without a way of removal. This condition results in disfigurement, physical discomfort and the loss of functionality in affected areas.[76] After biopsy of the tissue, it is usually placed in formalin and subsequently embedded with paraffin. It can then be sectioned and stained for evaluation by a pathologist.

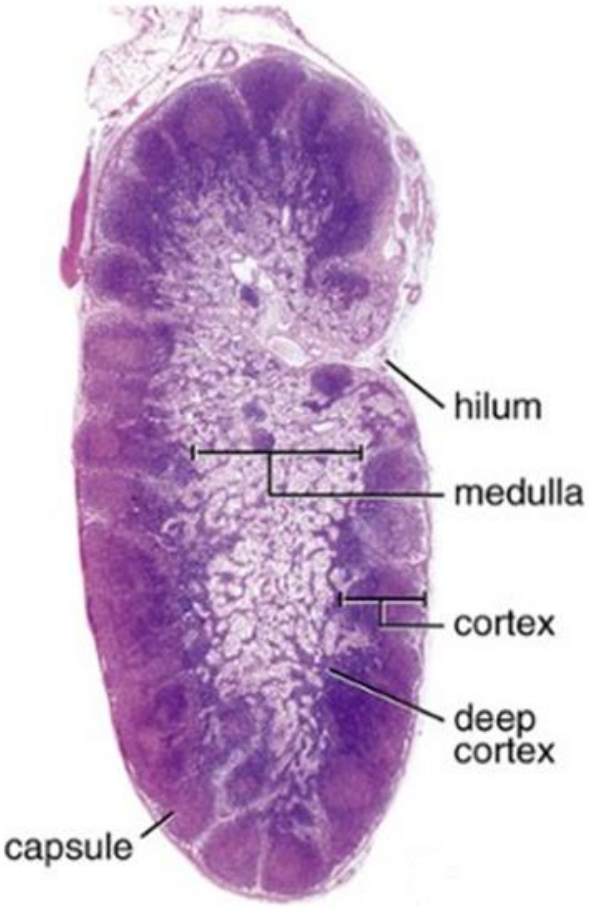

Figure 1.10- H&E stained lymph node section. Replicated from reference.[77]

Staining is done in order to clearly visualise cellular components in the section. Haematoxylin and Eosin (H&E) are often the stains of choice for histology. Haematoxylin dyes the cell's nuclear components blue and eosin dyes the cytoplasm of cells and connective tissue pink.[78] The cut surface of a normal lymph node is grey/pink in colour, has a soft texture and is homogenous. Whereas, the cut surface of a lymph node with neoplasia has a white colouration, a firm texture and is nodular.[79] Neoplasms are typically identified by increased cellular crowding and disorder to the tissue structure.[30] Analysis consists of assessment of the tissue's morphology, architecture and staining to form a diagnosis. Neoplastic cells exhibit cytological differences: increased nuclear material, inflammation, increased nuclear-to-cytoplasmic ratio, increased nuclear staining, abnormal nuclear shape and size, abnormal chromatin distribution, decreased differentiation and increased mitotic activity.[80] Figure 1.10 shows a microscopic image of an H&E stained lymph node section. The presence of

Hodgkin and Reed/Sternberg (HRS) cells (Figure 1.11) aid the differentiation of Hodgkin's lymphoma from non-Hodgkin's lymphoma; however, only 1% of these cells are present in tumour tissue.[81] Tindle states that the presence of a Reed-Sternberg cell alone does not identify Hodgkin's lymphoma but a diagnosis will not be made without them.[82]

Tissue can also be frozen rather than formalin-fixed but whilst metastatic disease is relatively simple for the pathologist to distinguish due to invasion of a different cell type, lymphoma is much harder to identify on frozen tissue sections.

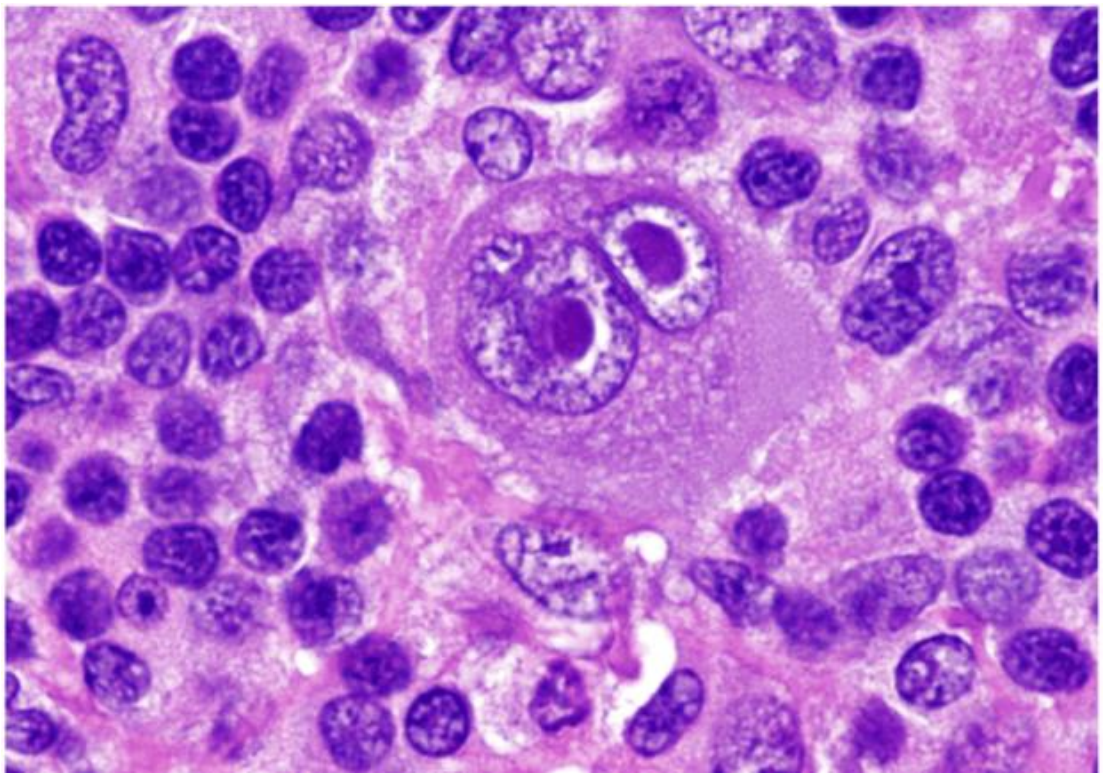

Figure 1.11- Classic binucleate Reed-Sternberg cell replicated from reference.[54]

This histopathological method of diagnosis requires a trained histopathologist, is invasive, due to the need of biopsy, have lengthy processing times, delays the start of treatment, and its findings may result in the need of subsequent surgical procedures. Histopathological examination can also be a highly subjective method of diagnosis[83] and it has been reported that diagnostic pathological discrepancies can occur in 1–43% of all specimens.[84] If the biopsy procedure fails to locate and remove tumorous tissue, normal tissue may be

analysed and result in false negatives. Treatments, which could potentially be life threatening, are chosen based on the outcome of histopathological diagnosis.[85] Tumours are heterogeneous and there are a wide range of cancer morphologies; thus, many tumours do not exhibit typical identifying features.[86] Tumours are poorly differentiated in adenocarcinomas and the microscopic appearance of the primary and metastatic cancers of different locations are similar to each other. Consequently histopathological examination is problematic and hence diagnosis is difficult.[87] It has been reported that there were major diagnostic disagreements between pathologists in 16.6% of Hodgkin's lymphoma cases and 27.3% of non-Hodgkin's lymphoma cases.[88] The analysis of sections by another pathologist to provide a second opinion is claimed to be too costly. However, Hahm *et al.*[5] argue that a second review would actually reduce the costs that misdiagnoses can cause. They go on to support this by stating that the cost of unnecessary surgery caused by misdiagnosis of a benign tumour as malignant, or the delayed treatment of a malignant tumour misclassified as benign, would be eliminated.

A WHO (World Health Organisation) classification could be assigned to 91% of non-Hodgkin lymphoma cases, however only 57.5% could be done so with high confidence.[89] WHO grading is typically more clinically reproducible than other classification methods.[89] The classification enables collaboration between the oncologists and pathologists and enables better understanding of lymphoid neoplasms and their treatment. Although pathologists primarily determine the classification of specimens, clinicians know the practical efficacy and outcome of classifications. WHO classification of lymphomas is based on cell lineage and morphology as either Hodgkin's lymphomas or non-Hodgkin's lymphomas: either B-cell neoplasms or T-cell neoplasms. T-cell and B-cell neoplasms can be further grouped into precursor neoplasms, representing the early stage of differentiation, and mature neoplasms, representing the tumours which are highly differentiated.[90]

Histopathological diagnosis involves the examination of only a portion of a lymph node. As a result, sampling errors can occur and lead to misdiagnosis. Figure 1.12 diagrammatically explains how sampling errors can occur.

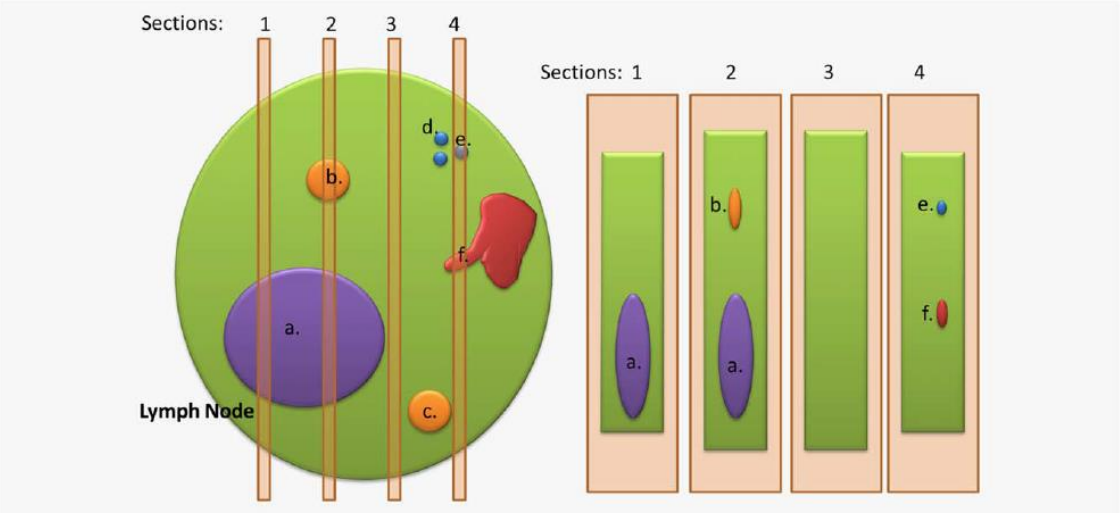

Figure 1.12- Diagram showing the results of sampling error as a result of tissue sectioning. Section 1 would result in a true positive diagnosis for a macrometastasis (>2 mm). Section 2 would result in a true positive for both a macrometastasis and a micrometastasis (<2 mm). Section 3 would result in a false negative diagnosis. Section 4 would result in a true positive diagnosis for isolated tumour cells (ITC) as well as a false negative diagnosis, where a macrometastasis appears to be a micrometastasis. Replicated from reference.[91]

1.2.6 Cytology

Cytological diagnosis is made as a result of the skilled interpretation of the microscopic appearance of cells. It is less reliable than histopathology for definitive diagnosis but is highly useful for detecting cellular differences in large population screening.[92] Fine needle aspiration cytology (FNAC) is an established method for the diagnosis of lymph node cancers. Although it cannot distinguish all types of neoplasms, it is a much safer procedure than open biopsy.[93] FNAC can be used in the diagnosis of common lymphomas such as large B cell lymphoma, small lymphocytic lymphoma and follicular lymphoma, which constitute approximately 70% of all lymphomas.[94] It is also a rapid and economical technique and is safer than core biopsies for superficial lesion analysis. It is also less invasive, due to a fine needle size, and hence reduces the trauma a patient can experience. The fine needle aspirates can be further examined using immunocytochemistry techniques. However, low sensitivities and specificities has led to questions concerning the further use of this technique diagnostically.[95] A wide range of FNAC sensitivities and specificities have been reported from 68–98% and 56–100% respectively.[96] Leenders *et al*. [97] calculated a sensitivity of just 24.7% for the identification of axillary metastases in breast

cancer using ultrasound guided FNAC, and concluded that the technique should not be relied upon alone for diagnosis.

1.2.7 Molecular Diagnosis

Molecular diagnosis offers the advantage of complete objectivity, precision and reproducibility. However, the current lack of accepted pathologically identifying biomarkers prevents the universal application of these techniques.[86] Cross *et al.* [98] stress the importance of identifying and detecting biomarkers for more accurate cancer diagnosis as the disease is highly complex. Usually molecular techniques employ a polymerase chain reaction (PCR) step using homogenised lymph node tissue or fine needle aspiration (FNA) samples. However, this is a destructive method and risk of contamination presents an issue for the reliability of results produced from this technique.[91]

Molecular diagnosis functions based on the fact that only specific genes are transcribed into messenger RNA (mRNA) in each cell. The mRNA in a cell dictates the proteins expressed and the amount present is dependent on the cell's lineage and differentiation stage, activity of intracellular regulatory pathways and the influence of extracellular stimuli. Firstly, mRNA must be extracted from the cell and copied, using enzymes, to produce fluorescent complimentary DNA probes which represent the expressed genes in the cell. The probes are incubated on the surface of a DNA microarray, which contain DNA spots for thousands of specific genes, to enable the probes to hybridise to the complimentary microarray spot. The extent of hybridisation can be quantified using a scanning fluorescent microscope. The expression of thousands of genes can be measured at one time with this technique. The diagnosis of cancers rely on the assessment of morphology, however differential patient response to treatment suggest that subcategories could exist within current clinical classifications. These subcategories could be recognised by the identification of varying gene expression in different cells.[99] Although gene profiling is an expensive technique it has been argued that it is cheaper than current methods in the long term; as accurate diagnosis leads to early and targeted treatment, resulting in better prognosis and increased survival rates which is more cost-effective.[100] However, DNA micro-array studies are limited to a few types of

cancers and have been investigated using a wide variety of methodologies, which can lead to the inability to compare the results of the studies to one another.[86]

Immunohistochemical staining involves the staining of tissue or cells with coloured/fluorescent dye tagged antibodies, which can identify cancer cells based on the alteration of protein expression when compared to normal cells [19]. However, technical failure or sample heterogeneity can result in lack of staining and thus misdiagnosis, as observed by Nielsen *et al.*[101] where the use of a single basal marker missed approximately half of basal-like tumours.

1.3 Optical Spectroscopy

Spectroscopy is the study of the interaction of electromagnetic radiation (Figure 1.13) with matter. Information about a sample's molecular structure can be gained, such as the symmetry, bond distances and angles. The chemical properties of a sample, such as electronic distribution and the bond strength, can also be acquired.[102] Techniques utilise the emission, absorption or scattering of this radiation, to gain the information regarding the molecules' identity. Emission is where a molecule drops from a state of high energy to a state of low energy, and thus emits the excess energy as a photon, such as fluorescence. Absorption of radiation by molecules can be monitored by measuring the intensities of transmitted light over a range of wavelengths in order to ascertain which wavelengths have been absorbed by the sample. Molecules also have the ability to scatter the light it interacts with.[103] This scattering can occur both elastically with no energy change, or inelastically, where energy is transferred between the light and molecule. It is these inelastic collisions that are detected in Raman spectroscopy.

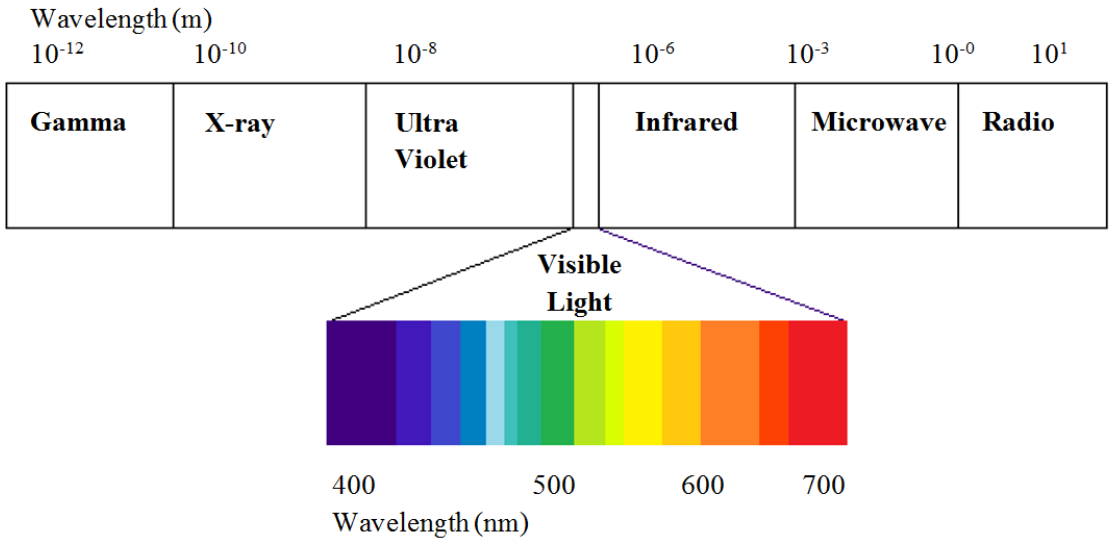

Figure 1.13- Electromagnetic spectrum.

1.3.1 Vibrational Spectroscopy

Spectroscopy utilises the interaction of light and matter to produce characteristic spectra unique to specific molecules. A molecule consists of atoms and the bonds that bind them together. These molecules consist of sets of coupled harmonic oscillators that vibrate when disturbed from a state of equilibrium.[104] A molecule exposed to a static electric field, associated with radiation, becomes distorted as the positively charged nuclei are attracted toward the negative pole and the negatively charged electrons are attracted toward the positive pole. The charge centre separates and an induced dipole moment occurs as a result, the molecule is polarised.[105] This dipole moment (μ) is proportional to the magnitude of the applied electric field (E) and the ease of molecular distortion[105] and can be represented by the following equation, where α is the polarisability:

$$\vec{\mu} = \alpha \vec{E} \qquad \text{Eq. 2.3.7}^{105}$$

Vibrational motions of a molecule involve a set of nuclei that are connected by forces that act between the nuclei. Strong forces occur between the nuclei of bonded atoms and provide resistance during the stretching mode of molecules. Weak forces occur between the nuclei of non-bonded atoms and provide resistance during bending modes that cause an increase or decrease to the angle. Displacement of the nuclei leads to complex motions of the molecule.[106] Molecular vibrations can occur by both stretching and bending of bonds (Figure 1.14). Stretching involves the variation of a bond length between two atoms, through lengthening and compression. Stretching can be symmetric, where it occurs in phase, or asymmetric, where it occurs out of phase. Bending is where an atom moves out of its regular plane, leading to a change in the angle of the bond.[107] The different modes of vibration can be seen in Figure 1.14.

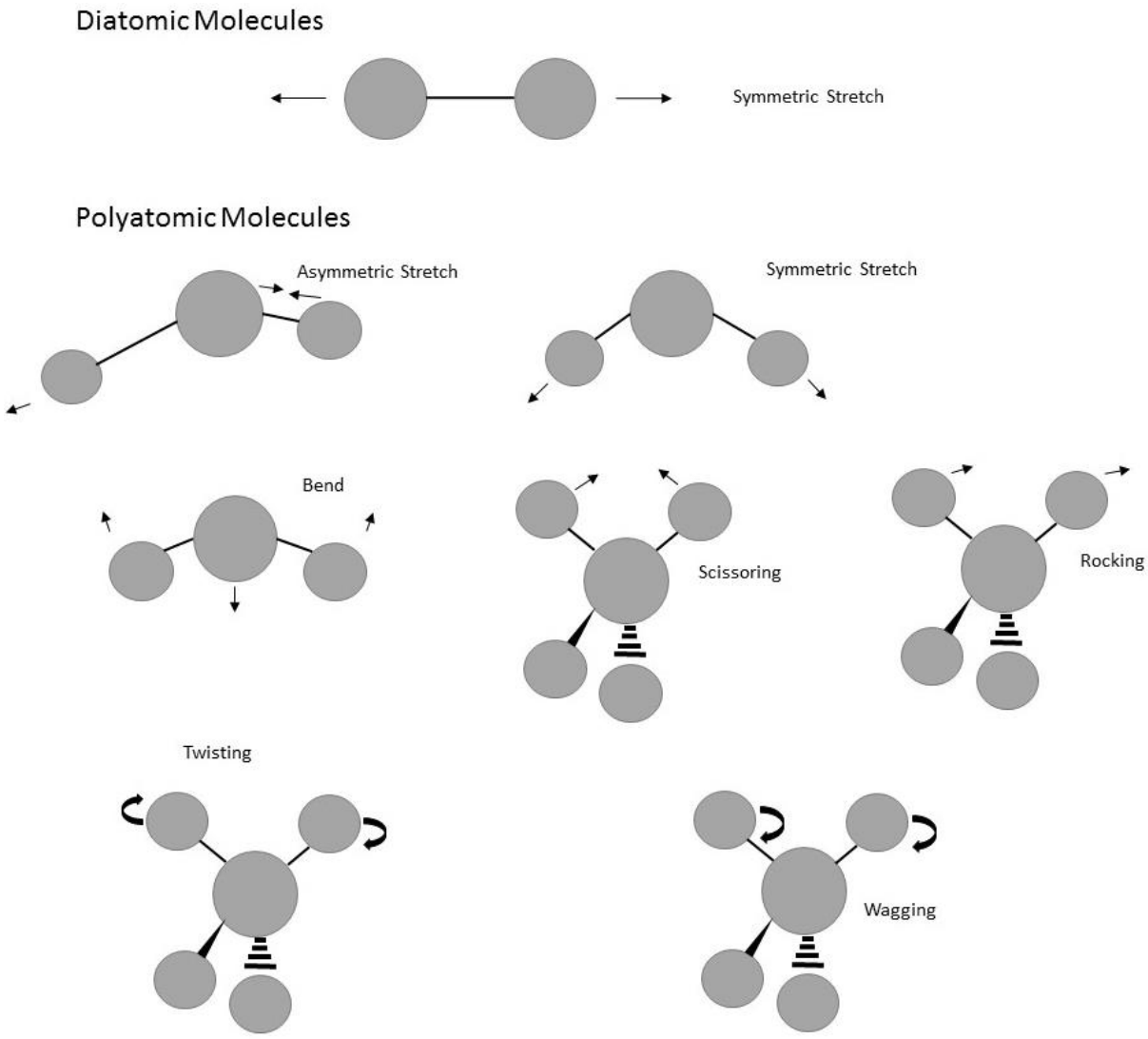

Figure 1.14- Vibrational modes.

Vibrating molecules can interact with radiation either by absorption, which is measured in infrared spectroscopy, or scattering, which is measured in Raman spectroscopy.[104] The frequencies of the Raman shifts measured in Raman spectroscopy correspond to the absorption frequencies measured in infrared spectroscopy. This is because each normal mode of vibration oscillates at a characteristic frequency and in turn either scatters or absorbs light at a frequency of either the difference or the sum of the incident radiation and the molecule's vibrational frequency, thus the energy exchange in both the absorption and scattering of light is specific to molecules (Figure 1.15).

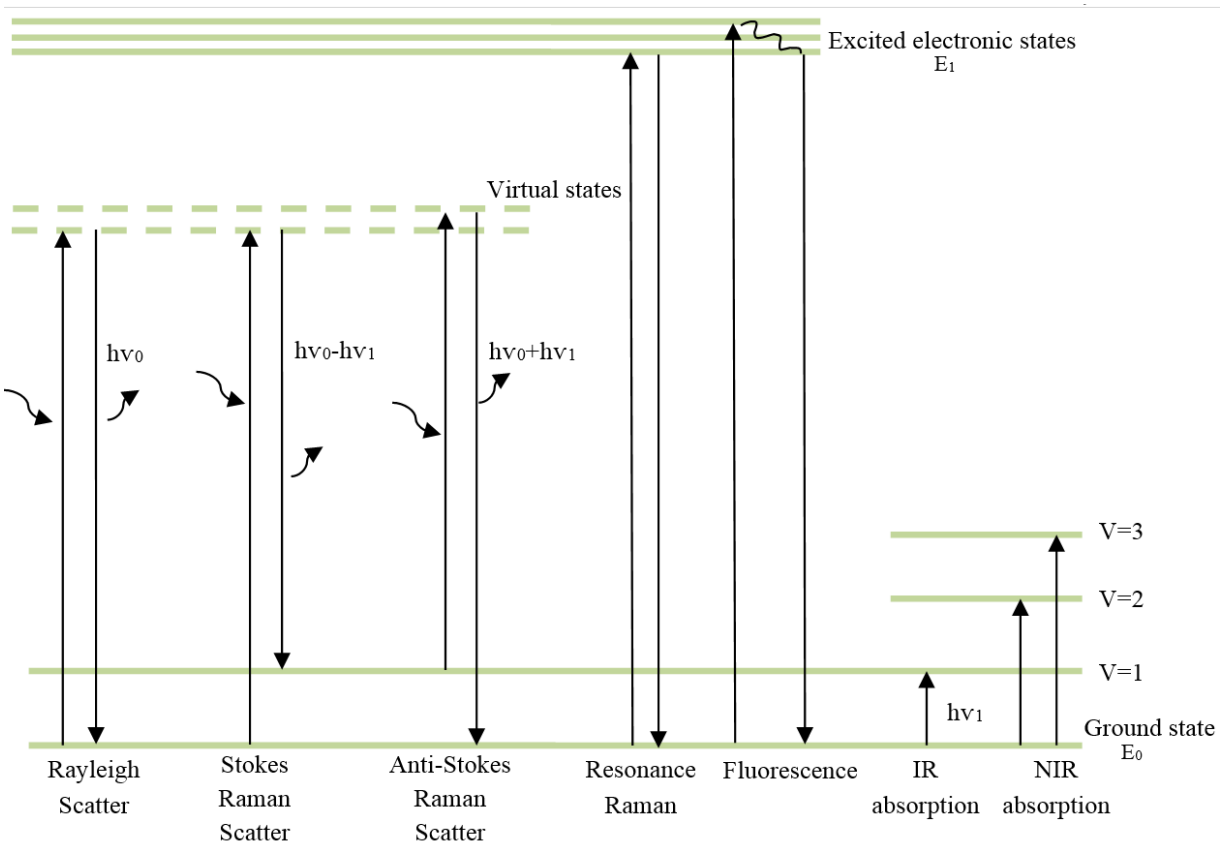

Figure 1.15- Energy transitions for different types of vibrational spectroscopy. The virtual state represents the temporary distortion of the electron distribution [108,109].

For the vibrational modes to be active, both techniques have different selection rules. Therefore, some vibrational modes are mutually exclusive, being only active in either Raman or infrared spectroscopy, and the band is expressed in only one of the technique's spectrum. The vibrational modes of molecules with a centre of symmetry are typically active in either Raman or infrared but not both.[110]

For a diatomic molecule there are three degrees of translational freedom and two degrees of rotational freedom. Only the symmetrical stretching vibrational mode is present in diatomic molecules. Polyatomic molecules containing (N) number of atoms will have 3N degrees of freedom (Table 1.4).[107]

Table 1.4- The degrees of freedom for linear and non-linear molecules. Reproduced from reference [107]

Degrees of freedom	Linear	Non-linear
Translational	3	3
Rotational	2	3
Vibrational	3N-5	3N-6

The degrees of freedom of a molecule depend on whether the molecule is of a linear or non-linear composition. The frequency of these vibrational modes is dependent on the masses of the atoms involved and the stiffness of the bond.[107] Hooke's law can be applied to the vibrational properties of atoms. Hooke's law states that the force (F) needed to move a solid from its equilibrium state by a specific distance (x), is related to the force constant (f), a measure of bond strength. In equation 2.3.1 the restoring force is shown by $-f$ (the force in the opposite direction to the elongation):

$$F = -f \cdot x \qquad \text{Eq. 2.3.1}$$

Newton's law states that force is also directly proportional to the mass (m) and its acceleration (the second derivative of the elongation with respect to time (t)):

$$F = m \cdot (d^2x/dt^2) \qquad \text{Eq. 2.3.2}$$

Combining these equations (2.3.1 and 2.3.2) results in:

$$m \cdot (d^2x/dt^2) = -f \cdot x \qquad \text{Eq. 2.3.3}$$

Solving this equation for x describes the motion of the atom as a harmonic oscillation where v is the vibrational frequency and φ the phase angle:

$$x = x_0 \cdot \cos(2\pi v t + \varphi) \qquad \text{Eq. 2.3.4}$$

The second derivative of distance (x) by time (t) is found to be:

$$d^2x/dt^2 = 4\pi^2 v^2 x_0 \cos(2\pi v t + \varphi) = -4\pi^2 v^2 x \qquad \text{Eq. 2.3.5}$$

Insertion into equation 2.3.3:

$$v = (1 \,/\, 2\pi)\sqrt{(f \,/\, m)} \qquad \text{Eq. 2.3.6}$$

The reduced mass (μ) is a combination of the atomic masses (m_1, m_2) that are involved in the vibrational mode, and is calculated by:

$$\mu = m_1 m_2 \,/\, m_1 + m_2 \qquad \text{Eq. 2.3.7}$$

The wavenumber of the vibrational mode can then be calculated, using the speed of light constant (c) as:

$$\bar{v} = (1 \,/\, 2\pi c)\sqrt{(f \,/\, \mu)} \qquad \text{Eq. 2.3.8}$$

Therefore the wavenumber of the vibrational mode is indicative of the bond stiffness reduced mass in order to understand the molecular structure.

Both Raman and Infrared spectroscopy have many advantages as analytical techniques. They are both rapid, non-destructive, objective, economical and their instrumentation is relatively small.[104]

1.3.2 Raman Spectroscopy

Raman spectroscopy is named after the Indian physicist Chandrasekhar Venkata Raman, who first observed and described the phenomenon of inelastic light scattering in 1928 and received the Nobel Prize for it in 1930.[111]

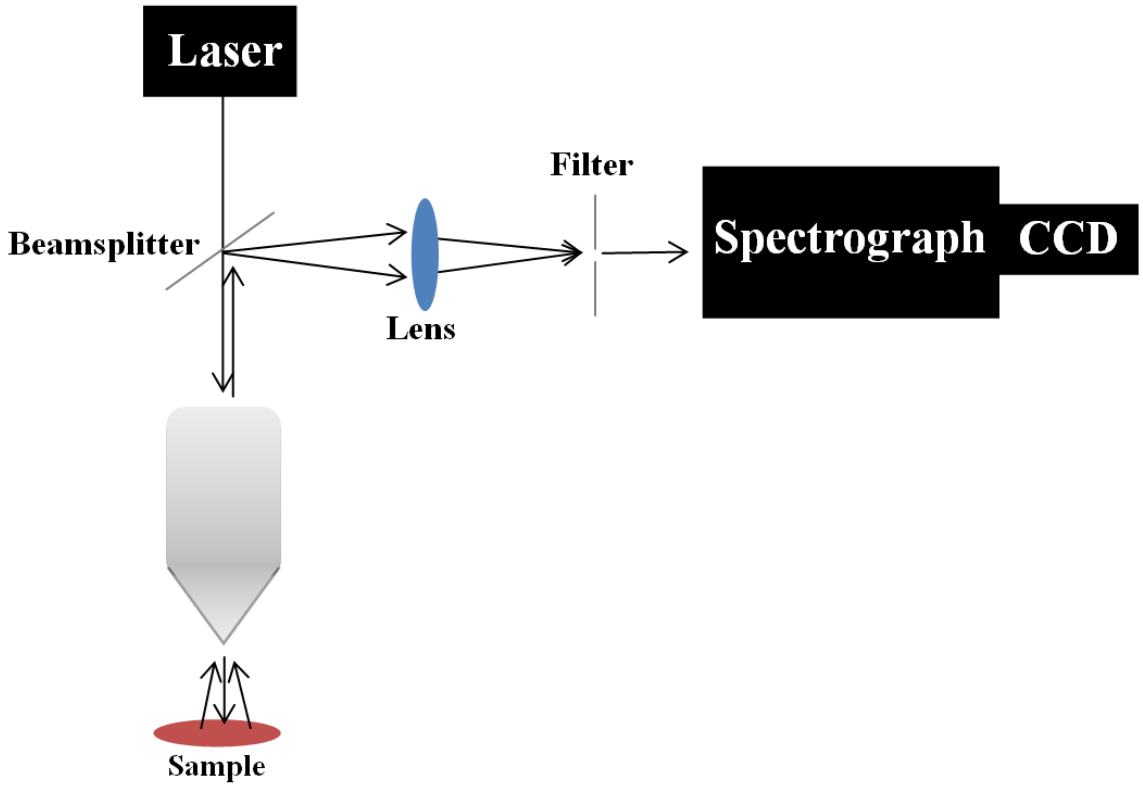

Figure 1.16- Schematic of a Raman Spectrometer coupled with an objective (silver).

Raman spectroscopy utilises a monochromatic and coherent light source (laser) for sample illumination, the interaction between the sample and the light results in photon scattering which is detected and measured. The instrumental schematic is shown in Figure 1.16. Scattering can occur either with or without an exchange in energy between the light and the molecule. When light illuminates molecules in a state of equilibrium, radiation of the same frequency as the incident light is scattered from the molecules. This form of elastic scatter is frequent and is referred to as Rayleigh scattering.[105] Raman scattering on the other hand, is inelastic and generally comprises less than 0.001% of the total

scattered light.[112] An energy exchange occurs between the molecules within a sample and the incident light in accordance with its vibrational frequency. The light is scattered with a different frequency, and thus wavelength, to that of the incident light. The wavelength change is directly related to specific normal modes of vibration. The energy change must be equal to the difference in energy between two of the allowed energy states of the molecule. When a molecule gains energy, photons will be scattered with hν-ΔE and the light will have a frequency of ν-ΔE/h. Conversely if a molecule loses energy, photons will be scattered with hν+ΔE and the light will be scattered with a frequency of ν+ΔE/h. Where hν is the photon energy, ν is the radiation frequency and h is Planck's constant (6.62×10^{-34} J.s).[105] Scattered light with a lower frequency to that of the incident light is called Stokes scattering. At room temperature the majority of molecules are found at ground state and thus, this form of Raman scattering is more common. Energy is transferred from the incident light to the molecule; the excited molecule does not return to its ground state from the virtual state, but to its vibrational state, and a lower energy photon is emitted (Figure 1.15). Scattered light of a higher frequency to that of the incident light is called anti-Stokes scattering. This is where the molecule is already at its vibrational state, due to prior excitement or thermal disturbance, and the molecule is excited to a virtual state. Energy is transferred to the incident light from the molecule, returning it to its ground state, and a photon of higher energy is emitted. Raman shift (cm^{-1}) is calculated based on the energy difference between the scattered light and the wavelength of the incident radiation.[113,114]

The ability of a molecule to scatter light is based on its polarisability. Polarisability represents the ability of an electric field to induce a dipole moment in a molecule. The electron clouds of molecules and atoms that are further away from the nucleus are easier to distort, thus have a stronger polarisability, than those that are closer to the nucleus. However, rotation and vibrations of a molecule can alter its polarisability, α, and can be explained by the following equation. Where, α_0 is the polarisability of a molecule at its equilibrium nuclear geometry, Δr is the distance from a molecule's equilibrium geometry, and $\left(\frac{\partial \alpha}{\partial r}\right)$ is the change in polarisability in relation to the change in the position during vibration or rotation.

$$\alpha = \alpha_0 + \left(\frac{\partial \alpha}{\partial r}\right) \Delta r \qquad \text{Eq. 2.3.10}$$

The sinusoidal behaviour of a rotating or vibrating molecule causes Δr to be written as:

$$\Delta r = r_{max} \cos(2\pi v t) \qquad \text{Eq. 2.3.11}$$

Where v is the frequency of the vibration, t is the time, and r_{max} is the maximum vibrational amplitude. The electric field, E, associated with light has sinusoidal behaviour and therefore is represented by the following equation. Where, v_{in} is the frequency of light and E_{max} is the maximum electric field frequency.

$$E = E_{max} cos(2\pi v_{in} t) \qquad \text{Eq. 2.3.12}$$

These equations can then be combined with equation 2.3.10 to produce:

$$\mu_{in} = \alpha_0 E_{\max} cos(2\pi v_{in} t) + E_{max} r_{max} \left(\frac{\partial \alpha}{\partial r}\right) cos(2\pi v t) \cos(2\pi v_{in} t) \qquad \text{Eq. 2.3.13}$$

The following equation describes the relationship for the product of two cosines:

$$\cos a \times \cos b = \frac{1}{2}(\cos(a+b) + \cos(a-b)) \qquad \text{Eq. 2.3.14}$$

From the previous equation (2.3.13) it can be seen that $2\pi v_{in} t$ is represented by 'a' and that $2\pi v t$ is represented by 'b' in equation 2.3.14, which results in the following equation:

$$\mu_{in} = \alpha_0 E_{max} \cos(2\pi v_{in} t) + \frac{E_{max} r_{max}}{2} \left(\frac{\partial \alpha}{\partial r}\right) \left(\cos\left(2\pi t (v_{in} + v)\right) + \cos\left(2\pi t (v_{in} - v)\right)\right) \qquad \text{Eq. 2.3.15}$$

The first term in the equation describes Rayleigh scattering, as both the incident and scattered light frequencies are the same as one another. The second term describes Raman anti-Stokes scattering as the scattered light has a higher frequency and lower wavelength than that of the incident light. Thus, the third term describes Raman Stokes scattering as the scattered light has a lower frequency and higher wavelength than that of the incident light. If the change in polarisability $\left(\frac{\partial \alpha}{\partial r}\right)$, is zero, the second and third term will equal zero and there will be no Raman scattering. Thus, the selection rule for vibrational modes to be Raman active, is that there must be a change in polarizability.[115] A change in a molecule's polarisability is reflected in either the magnitude or direction of the 3-dimensional polarisability ellipsoid.[105]

The scattered light is then detected and in Raman spectroscopy this is done using charge coupled device (CCD) detectors, primarily composed of pure silicon. The detector's quantum efficiency (QE) is dependent on silicon's absorption ability. At wavelengths outside 3500-8000 Å photons will pass through the silicon, be reflected off the surface or absorb into only the surface layers of the silicon.[116] The quantum efficiency is calculated by: QE = Number of photons emitted/Number of photons absorbed.[117]

Silicon atoms covalently bond to one another and their electrons are not detectable when they are bound to the structure. If the structure receives energy that is greater than the band gap (photons that have wavelengths shorter than 10 m) an electron hole pair is created. The electron can then be promoted to a conduction band and detected. Thermal energy can also produce free electrons which result in signal, not due to photons, called noise. These free electrons are collected and stored on a silicon chip defined by electric field regions, called 'pixels', aligned in rows and columns. The electrons are read-out off the chip in order to produce a digital signal. A shift register runs horizontally below the pixel matrix to enable the signal to be read off the chip through shifting the entire charge pattern one row at a time to the shift register using electrodes. Vertical electrodes are then used to transfer the charge pattern across the shift register one element at a time into the on-chip output amplifier. The signal then is

transferred to the analogue-to-digital (A/D) converter which produces a 16 bit binary code of each charge packet. Binning can be used to allow faster readout and the production of reduced noise spectra due to the combining of two or more pixels on the CCD chip.[118] The use of a CCD detector allows multiple pixels to be detected simultaneously as opposed to a single detector which would require a rotatin grating and the acquisition of each data point separately. Therefore CCD detectors speed up spectral collection by the channels in the CCD on the wavelength axis

1.3.3 Raman Spectra

Raman spectra are plotted as Raman shift against the intensity of the scatter. The Raman shift is typically presented as the difference in wavenumber between the incident and scattered light, and is independent of the incident wavelength. Therefore, it is characteristic to specific molecular vibrations and peak positions will remain the same for the same specimens regardless of the wavelength of the incident laser chosen to take measurements. The intensity of the scatter is directly proportional to the concentration of the responsible molecules within a specimen.[119] Therefore, Raman spectroscopy is both a qualitative and quantitative technique that produces a characteristic spectral fingerprint of samples.

The Raman spectra can be considerably affected by artefacts and non-sample contributions. Gaussian noise can occur as a result of fluorescence of the sample or thermal fluctuations on the charge coupled device (CCD), and can result in an unwanted intense spectral background. The most significant source of random noise is shot/Poisson noise. It occurs as a result of the random probability of detecting photons at specific wavelengths. Shot noise is proportional to the square root of the total counts measured. Therefore, for a spectrum of n number of counts, the associated shot noise will be $n^{1/2}$. Due to this square-root relationship, although the variation in counts will increase as the total number of counts increases, the relative error will decrease. The signal-to-noise (S/N) ratio is a measurement for the relative error and represents the total number of counts divided by the noise associated counts.[120] Increasing either the

acquisition time or the number of accumulations will reduce the overall noise contribution in the resulting spectrum. Thermal/dark noise is detector associated and arises due to the generation of spontaneous electron-hole pairs, even in the absence of photons. This noise occurs as a result of CCD temperature and can therefore be reduced by cooling the detector. Readout noise occurs as a result of the conversion of the analogue signal to digital. It originates from the electric components, such as amplifiers, and the analogue-to-digital convertor, and it is not proportional to the measurement time.[121] Hardware binning can be carried out prior to signal readout by the preamplifier, it combines the pixel charges on the CCD and thus improves the S/N ratio. Binning improves S/N ratio of readout noise limited data linearly with the number of pixels co-added, and improves S/N ratio of shot noise limited data by the square-root of the number of combined pixels.[122] Noise contributions should be reduced as much as possible experimentally.

Cosmic spikes can also arise in spectra, caused by high energy particles (cosmic rays) hitting the CCD. These artefacts can be eliminated, most simply, during spectral measurements by taking multiple accumulations and also applying a median filter.[123]

More intense backgrounds can be caused by the substrate chosen to support the sample. Fullwood *et al*.[124] showed that calcium fluoride (CaF_2) windows (Crystran, UK) result in lower spectral backgrounds when compared with low-emissivity Mirr IR (Kevley technologies) and spectrosil quartz slides. Lower spectral contributions from substrates reduces the rquirement of background correction techniques. Near-IR radiation can have a penetration depth of 1 mm.[125] Therefore with samples which are less than this, the substrate will be probed and scatter light, thus it is important to have a substrate which only minimally contributes to the resulting spectrum.

1.3.4 Raman Spectroscopy for Biological Materials

Raman Spectroscopy is a non-destructive, minimally-invasive, rapid and relatively low cost technique that can provide objective and reproducible diagnostic information about biological samples based on their biochemical

constituents. As a tissue develops from a normal to neoplastic state, biochemical changes occur. DNA concentrations increase and the concentrations of other components in the cell change.[126] These biochemical changes should be reflected in its Raman spectra and thus enable molecular information to be obtained and a pathological diagnosis made of tissue samples. Biological tissue is highly heterogeneous and consists of a number of components, including blood, lymph, fibrous tissue and parenchymal tissue, some of which contribute towards noise and fluorescence in Raman spectra when excited.[127] Higher frequency lasers increase the intensity of the Raman scatter, $I \propto \lambda^{-4}$, however they also result in fluorescence and phosphorescence caused by the induction of electronic transitions which result in light re-emission.[128] Biological chromophores include haeme, carotenoids, bilin and chloroplasts.[129] Intense spectral backgrounds are often attributed to fluorescence as a result of biomolecules within the sample absorbing at the excitation wavelength. Although, short-wavelength lasers are more likely to be responsible for the production of molecule fluorescence within biological specimens and for photolytic degradation of a sample.[130] Lasers with longer wavelengths, such as in the near-infrared region, produce less of these phenomena and can often be used at much higher output powers. Intense spectral backgrounds can still occasionally be measured at longer wavelengths, however these are not the result of sample fluorescence. The excitation wavelengths of biological chromophores are generally below 500 nm, and although haemoglobin fluoresces at longer wavelengths, little occurs above 600 nm.[131,132] Lasers with wavelengths below 270 nm can also be used to reduce fluorescent contributions in spectra due to the wavelength shift of the fluorescence at these wavelengths; however, the penetration depth is limited to only a few micrometres.[133] UV and blue light is readily absorbed by molecules within biological samples, such as haemolglobin, leading to fluorescence and thus has a much shorter penetration depth of <100 μm. Near infrared light is absorbed less by haem molecules and therefore has a deeper penetration depth of around 1000 μm.[134] Puppels *et al*.[135] studied the presence of carotenoids in lymphocytes in order to investigate previously reported inverse relationships between cancer and the concentration of carotenoids in blood plasma.[135,136] Using Raman spectroscopy they observed that carotenoids generally accumulate in the gall body of lymphocytes and suggested their use in identifying individuals at risk of developing cancer.

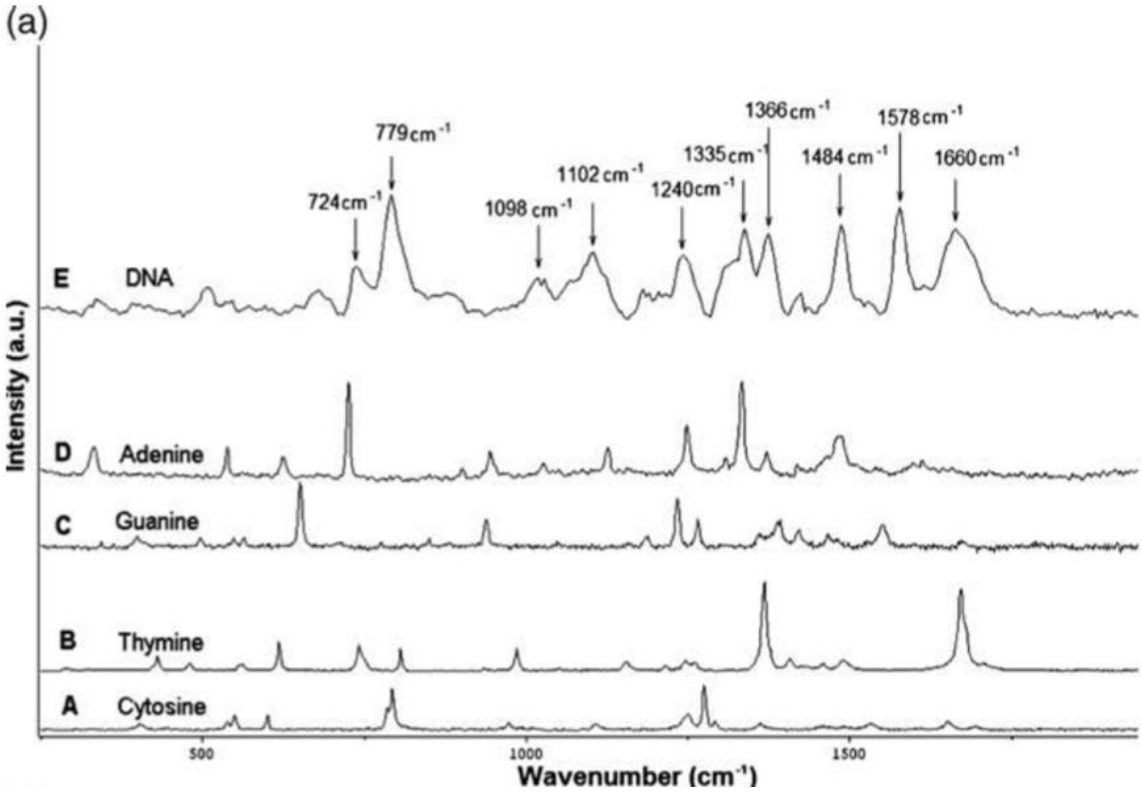

Figure 1.17- Spectra of DNA and its nitrogenous bases. Replicated from reference [137].

Huang *et al*.[138] studied Raman spectroscopy of various organs in murine (rodent) models using a catheter for probing the surface of the tissue with a 785 nm laser. The catheter consisted of low hydroxide fibres to deliver the incident radiation, one fibre of 200 μm diameter, and collect the scattered light, 27 fibres of 100 μm diameter. They concluded that Raman spectroscopy is a useful technique for disease diagnosis and observed that the spectra taken from tissue types that are alike are also similar to each other. For example, the spectra measured from bone and teeth tissue are similar, attributed to their high calcium content. Spectral measurements taken from the stomach, small intestine, colon and bladder were also found to be similar due to their structural similarities and high muscle compositions. They found that the Raman spectra of biological samples primarily comprised lipid and protein peaks, however, the intensities and ratios of these peaks altered depending on organ. They also noted that unique peaks are also present in spectra, observed only in specific tissue types, such as the brain, lungs and bone.

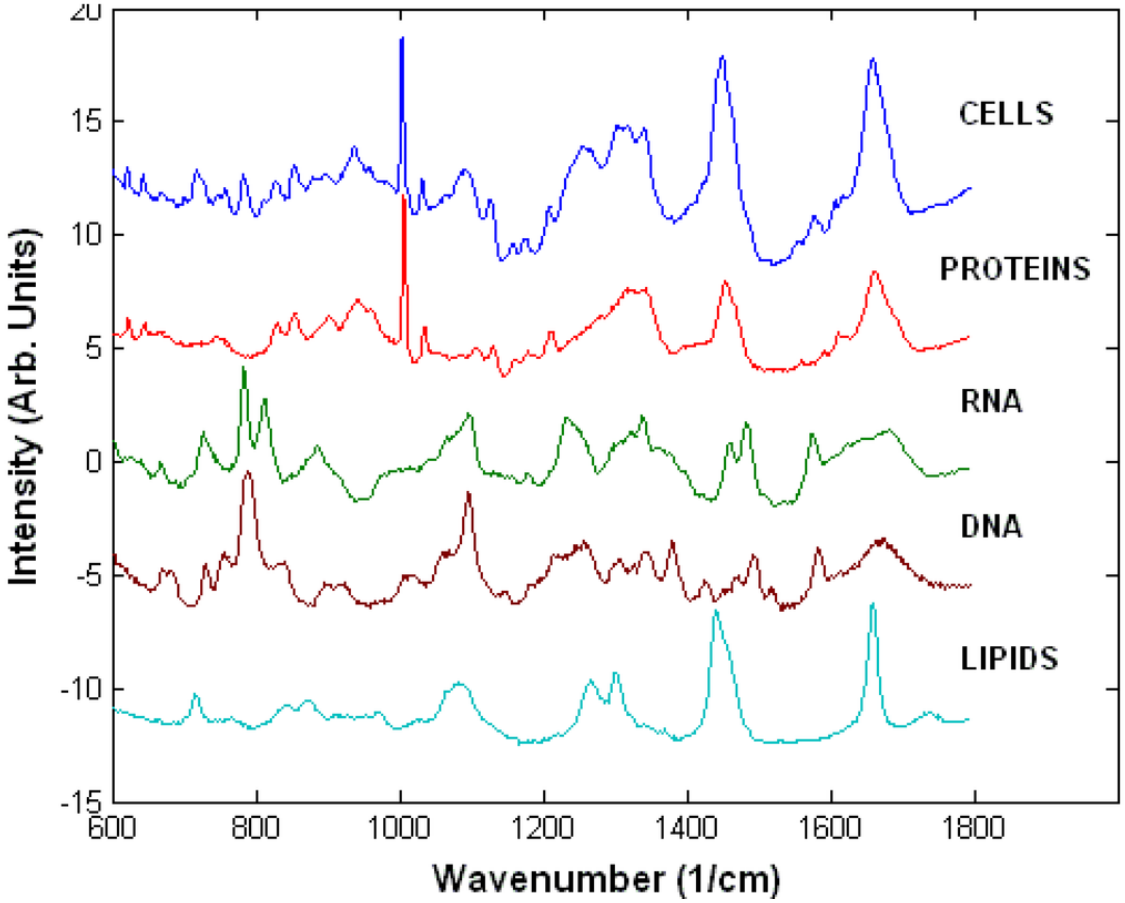

Figure 1.18- Raman spectra of human cells and constituents. Reproduced from reference.[139]

Raman spectroscopy utilises radiation of the electromagnetic spectrum that is non-ionising (when using visible or IR excitation). The -OH vibrational modes of water molecules have low polarisability and generate weak Raman signals, and thus do not substantially contribute to resultant spectra. As a result, Raman spectroscopy can analyse fresh tissue, with no prior sample preparation, both *ex vivo* and *in vivo* without destroying biological samples. The use of fibre optic probes can aid *in vivo* Raman analysis of biological tissues and enable real-time diagnosis, when inserted into a needle or endoscope.[140] Optic probes have fibres that deliver the incident radiation to the sample and fibres that collect the scattered light, carrying light through total internal reflection, and take it back to the spectrograph, enabling remote detection.[141]

Motz *et al*.[142] states that the fibre optic probe can guide surgery through identifying malignant tumours from surrounding non-cancerous tissue to delineate tumour margins and thus aid complete resection. Fibre optic probes are

limited by anatomical location and in the case of a needle probe, consideration must be taken regarding the gauge of the needle. A compromise must be made between an acceptable needle size in order to reach the target organ and a patient's discomfort whilst being efficient at delivering and collecting radiation in order to produce spectra with good signal-to-noise. They observed a 4x increase in background using a 785 nm laser compared to an 830 nm laser. They note that most *in vivo* applications are carried out using commercially available probes and are restricted to accessible organs such as the skin. Their study shows the ability to gather appropriate measurements *in vitro* on artery and breast tissue with their fibre optic probe design with an acquisition time of only one second which supports clinical application.

Another consideration is the safe exposure limit of this radiation. Although non-ionising, thermal heating and burning can occur as a result of irradiation. The laser wavelength, power and spot size all influence this. The international commission on non-ionising radiation protection[143] state that for thermal damage to occur the temperature of the skin must be raised to 45 ^{o}C. However for shorter exposure times a temperature of 55 ^{o}C must be achieved for just 10 seconds and 69 ^{o}C for just 1 millisecond. They suggest the limit H^{EL} for skin exposure at wavelengths less than 3000 nm and under 10 seconds, can be represented by the following equation:

$$H^{EL} = 2 * t^{0.25} * 10^{4} Jm^{-2} \qquad \text{Eq. 2.3.16}$$

Shim and Wilson[144,145] measured the spectra of different probe fibres, silica and sapphire. They established that whilst the spectra of low-OH silica and sapphire are identical, sapphire fibres have a strong background and thus it was determined that low-OH silica-silica fibres are the best choice for fibre optic probes in the near-IR region of radiation.

Day and Stone[146] describe their development of the incorporation of a fibre-optic Raman probe with a hypodermic needle (Figure 1.19). Their design is a low-cost and simple needle probe for subcutaneous use and is able to deliver and collect enough signal to identify and differentiate cancerous tissue from non-

cancerous with acquisition times of just 10-30 seconds. The probe incorporates just two fibres, one that delivers the light and one that collects the scattered light. They suggested that improvement of the design can be made through the use of multiple collection fibres to increase the signal detected. Iping-Petterson *et al*.[147] have developed this needle probe further to incorporate one delivery fibre surrounded by six collection fibres and show the improved signal collection than the two fibre probe, and improves the signal-to-noise ratio of collected spectra by 2.5 times. Immediate diagnosis can lead to early treatment, and Raman spectroscopy could aid biopsy, targeting the identification of suspicious areas or interrogation of tumour margins during surgery. This will reduce the risk of tumour recurrence and subsequent operations.[148] This method of diagnosis has economic benefits as it reduces the need for hospital resources and personnel.[3]

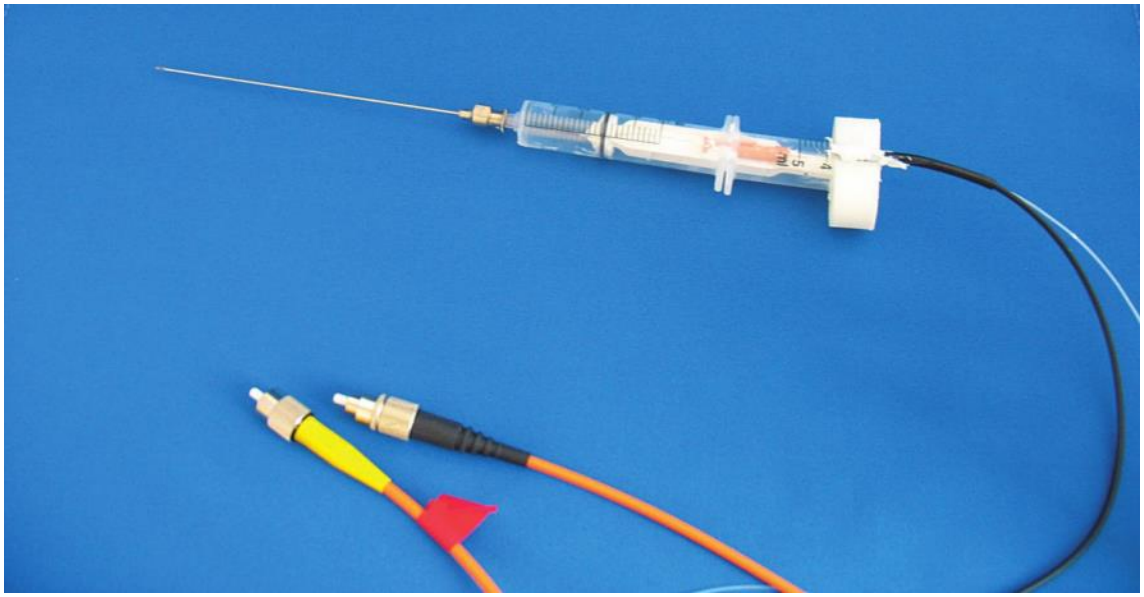

Figure 1.19- Day and Stone's design of a fibre optic Rama probe incorporated into a hypodermic needle, consisting of delivery and collection fibres. Reproduced from.[146]

Standard Raman spectroscopy is limited by a penetration depth of a few-hundred microns, thus restricting laser probing to the surface or near-surface area of samples.[149] The development of spatially offset Raman spectroscopy (SORS) enables spectral measurements to be taken from areas deep into the sample. The technique is based on the principle of photon migration. Scattered photons are collected from surface areas on samples that are laterally offset at varying distances from the incidence point of the excitation laser. A range of

spectra are generated with different contributions in peak intensities, which are relative to the surface and subsurface layers. As collection becomes further away from the point of laser excitation, the spectral contributions that relate to the top layer of the sample decrease faster than those that relate to the deeper layers. This is due to the probability of greater lateral migration of photons which are generated deeper in the sample before emission from the sample surface. Through processing of the dataset using multivariate analysis, pure spectra representing the individual layers of the sample can be produced.[150] Deep Raman enables measurements to be taken from areas several millimetres deep into tissue samples, thus aiding in vivo application.[151]

The detection of individual molecules has been demonstrated through the identification of DNA bases using surface enhanced Raman spectroscopy (SERS). This would enable the use of this technique as a label-free biomolecule method of detection.[152]

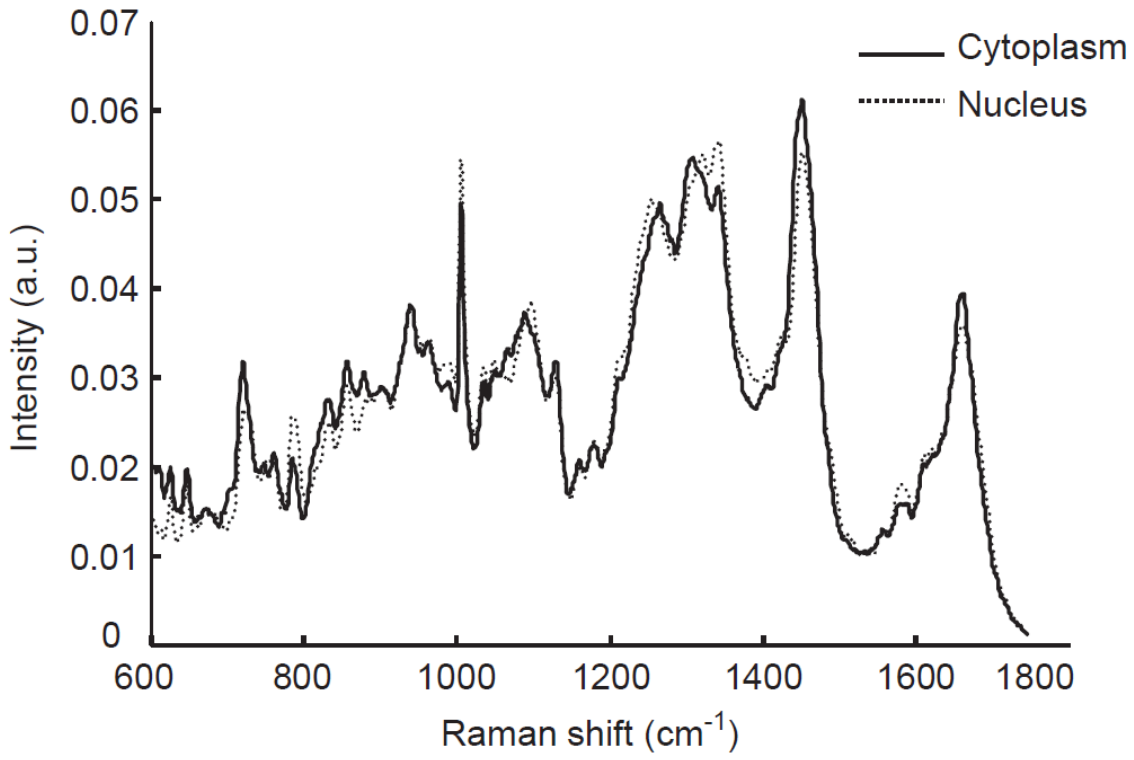

Figure 1.20- Spectra taken from the cytoplasm and the nucleus of a cell. Reproduced from reference.[153]

Figure 1.20 shows the ability of Raman spectroscopy to produce spectra at the sub-cellular level and shows the biochemical differences between the cytoplasm and nucleus of the cell that are reflected in their spectra. Matthaus *et al.*[154] observed that, with the use of multivariate analysis, Raman microspectroscopy could distinguish between components of the cell such as the nucleolus, nucleus, cytoplasm and mitochondria based on their biochemical composition. The average size of a human cell is 10-20 μm, as a result of the laser spot sizes for Raman spectroscopy, the nucleus of a cell is often not probed alone and spectra will contain signal from other cellular components. However, lymphocytes have a large nuclear to cell ratio and through histochemical staining it has been observed that the nucleus accounts for approximately 80% of lymphocyte cells. Therefore, it is likely that the nucleus is predominantly probed when investigating these cells.[155]

1.3.5 Infrared Spectroscopy

Infrared (IR) Spectroscopy is the study of the absorption of radiation, in the infrared region of the electromagnetic spectrum, by molecules within a sample. For absorption to occur, the frequency of the incident light must be the same as that of one of the normal modes of vibration. The absorbed wavelengths are determined in order to create a spectrum, which peaks correspond to the specific normal mode of vibration.[104] An infrared spectrum is a plot of the quantity of the light absorbed, represented by absorbance units, on the y-axis against the wavenumbers in cm^{-1} on the x-axis. The Beer-Lambert law, $A=\varepsilon bc$, states that absorbance A is directly proportional to molar absorptivity ε (in L mol^{-1} cm^{-1}), the path length of the sample b (in cm), and the concentration of the sample c (in mol L^{-1}). The absorbed light is calculated through measurement of light transmission using the equation:

$$A = \log(I_0 / I) \qquad \text{Eq. 2.3.17}$$

Where I_0 is the incident radiation and I is the transmitted radiation. Interpretation of IR spectra enables information about the sample under analysis to be gained. Molecules within a sample can be identified due to the peak

positions that correspond to specific wavenumbers absorbed. The amount of light absorbed is directly related to the concentration of the specific responsible molecules.[156]

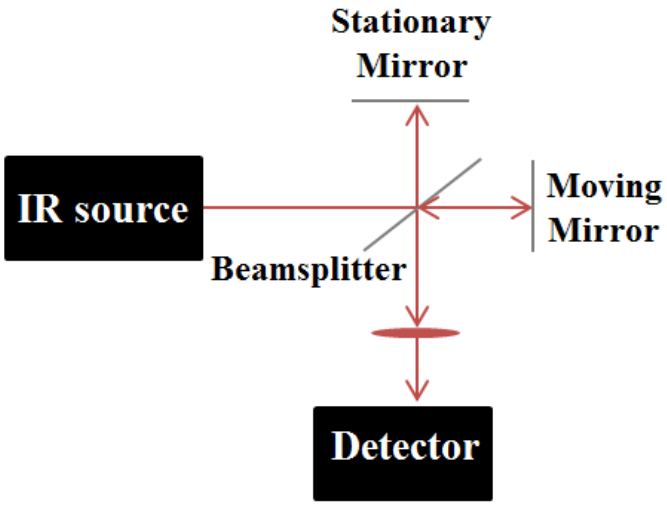

Figure 1.21- Schematic of an FTIR spectrometer.

Infrared light can be either absorbed or transmitted through a sample. IR spectroscopy can be used to measure the transmitted light, to determine what wavelengths were absorbed by the sample. This results in a spectral fingerprint of a sample that can indicate its molecular composition. Molecules can become excited and absorb light if there is a change in its dipole moment. The frequency of light absorbed is dependent on the energy gap between the ground state and excited state of its vibrational levels. The number of absorption peaks seen in the resulting spectra is based on the number of degrees of freedom for the molecule. Hence the molecular constituency and structure can be determined for a sample.[107] Fourier-transform IR (FTIR) spectrometers are currently the most widely used IR spectrometer, replacing the dispersive instrumentation that first became available in the 1940s.[107] The main reason for the development of FTIR was the need for faster spectral acquisitions. Dispersive instrumentation employ prisms or gratings to split the IR radiation into individual wavelengths, the detector then measures the amount of transmittance at each frequency. FTIR instrumentation utilise interferometers where individual spectral scans take only a second through simultaneous detection of high resolution spectral data.[157] An interferometer involves a beam splitter that transmits 50% of the incoming IR

radiation, emitted from a blackbody source, and reflects the other 50%. This therefore splits the radiation into two beams where one is reflected off a static mirror and the other is reflected off a mirror that vibrates a few millimetres to/from the beam splitter.[158] The two beams then recombine back at the beam splitter and exit as an encoded signal. The signal is an interferogram whereby every data point has information pertaining to every individual IR frequency of the IR source. The beam then reaches the sample where specific frequencies of energy can be absorbed. A detector collects the interferogram signal where it is digitised and sent to a computer for production and display of a frequency spectrum. This is done through the application of a mathematical technique called Fourier transformation. A background spectrum must be collected for removal of the part of the spectrum which was not contributed by the sample being analysed (typically consisting of water and CO_2.[157]

Fourier-transform Michelson type interferometers used in Infrared spectrometer instrumentation exploit the Fellgett, Jacquinot and Connes advantage. The Fellgett advantage is improved signal to noise ratio as a result of simultaneous detector measurement. Jacquinot advantage also offers increased signal to noise ratio but as a result of high optical throughput.[159] Connes advantage is an improvement in frequency accuracy in the FTIR instrumentation.[160]

As well as decreased acquisition time, there are other advantages to using FTIR instrumentation. Almost any sample in any state, such as gas, liquids and solids, can be analysed to generate a molecular spectral fingerprint, and the instrumentation is relatively cheap and simple to operate. FTIR spectrometers are also highly sensitive, detection of down to picograms of molecular constituents, and a good signal-to-noise ratio, which is a measure of peak size to assess spectral quality. However, there are disadvantages to this technique. Complex mixtures can result in spectra that are complicated and difficult to interpret. Molecules that do not possess IR active molecular vibrations cannot be detected by this technique, such as, homonuclear diatomic molecules. IR spectroscopy is highly affected by water in a sample and results in broad and intense OH bands in resultant spectra that can mask other sample peaks.[156]

Infrared focal plane array (FPA) detectors consist of an array of pixels, which are photon-sensitive in the infrared region, on the focal plane of a lens. The detector surface is a photovoltaic panel that detects photons and produces a voltage. The output voltage can then be digitised for computer read-out to construct an image of the sample. Each pixel of the detector represents an entire separate spectrum and therefore the number of spectra is determined by pixel resolution. FPA detectors allow rapid simultaneous detection of thousands of spectra, thus speeding up acquisition time.[161]

1.4 Data Analysis

1.4.1 Spectral Pre-processing

Pre-processing methods are often applied to spectra prior to multivariate analysis to remove unwanted non-sample contributions.[162]

Spectra can be green glass corrected to reduce the instrumental contributions in spectra. This involves taking a measurement of green glass on the instrument and subtracting the spectra away from subsequent sample signal in order to remove spectral contributions which result from instrumental fluorescence.

Baseline subtraction techniques are often applied to the data to remove remaining unwanted spectral baseline. A polynomial fit is a method of baseline correction where non-constant, low frequency baselines are adjusted for. Polynomials are calculated to fit a line to the data, preferably under the spectral baseline, which can then be subtracted from the data in order to decrease the intensity of the background. Afseth *et al*.[163] reported that baseline correction methods should be carried out on the data prior to normalisation. They state that derivatisation methods can be used as baseline correction techniques; however, they can lower signal-to-noise ratio. Although polynomial fits are not strongly reproducible in their baseline subtraction, they retain the same signal-to-noise ratio as the original data.

Spectra are often noisy and are thus frequently subjected to a degree of smoothing. Noise is usually high frequency and fluctuates randomly from one pixel to the next. Whereas, sample signal is usually of much lower frequency and gradually varies from one pixel to the next. This assumption is what smoothing techniques function on.[120] Smoothing can often be based on methods of local averaging that is applied to the data to de-noise each spectrum.[164] However, the ultimate objective of smoothing is to remove the noise without altering the relevant spectral features present. Savitzky-Gloay smoothing incorporates fitting an approximate polynomial across a selected window and taking the centre point, it then moves along one and repeats. Savitzky-Golay smoothing is less efficient

at reducing noise than compared to moving-window averaging methods, but is better at retaining the original signal shape.

Cosmic ray removal is also commonly carried out on spectra to eliminate peaks caused by cosmic rays. Cosmic rays are caused by photons in space rather than from the sample scattering laser light. Cosmic spikes are detected at random and are not related in time or space. Therefore, they can be eliminated from spectra through multiple accumulations or through pre-processing whereby the peak caused by cosmic rays are recognised and removed, either through removal of the whole spectra or by replacing the spectra between the affected wavenumbers.[165] On a Raman map median filtering can be applied for removing these peaks, through calculating the median spectra of a specified number of pixels and replacing the centre pixel with the median.

Biological samples are heterogeneous and the environmental and instrumental settings, such as the laser power, are changeable. Thus, normalisation is typically carried out on all the spectral set to mathematically account for this inconsistency, to adjust and reduce the non-sample attributed intensities. This is achieved through calculating the sum of squared averages of the data and then dividing the dataset by the square root of that sum.

Mean centring is often carried out on spectral data in order to allow interpretation of the intercept and calculate predicted values. The mean of the dataset is subtracted so it equals 0 in order to focus on the variation between the dataset rather than the variation between 0 and the mean.

1.4.2 Multivariate Analysis

Multivariate analysis is needed in order to overcome the limitations of univariate analysis, where each feature is taken as independent and the relationships between the groups are not considered.

Multivariate analysis is carried out to statistically assess variation within datasets[166] and to evaluate the level of discrimination between different groups within the dataset.

Clustering techniques such as K-means can be useful to assess similarities within the dataset. It is an unsupervised learning method where data are iteratively assigned to a specified number of groups (K) based on the Euclidean distance from centroids. These centroids are initially randomly assigned, the data are subsequently assigned based on the centroid it shares most similarity to, and then the centroids are eventually reassigned based on the average of the dataset within its group.[167] The number of clusters can be chosen based on the number of grouped may be expected within a dataset.

Principal component analysis (PCA) is often used for the identification of data points that are responsible for spectral differences and for analysing the level of discrimination between the spectral groups. Principal component analysis (PCA) is an unsupervised method of identifying patterns in data, in order to highlight the similarities and differences between the dataset.[168] The primary objective of employing PCA is to reduce the dimensions of a dataset, with a large number of interrelated variables, whilst preserving as much variation between the dataset as possible. This process is achieved through transformation of the principal components onto a new set of uncorrelated variables which are ordered to contain the majority of variation of the dataset in the first few components.[169] Dimensionality reduction offers advantages to the analysis; such as, a reduction in the required computational power, a decrease in the level of white noise, and the evasion of over fitting. Over fitting occurs when the algorithms include noise or spectral information that does not describe the difference in classification, which increases the accuracy of classification of the training dataset, however reduces the classification accuracy of an unknown test set.[165] Analysis of variance (ANOVA) can be used to determine the most significant components for delineating between the different groups of a dataset. ANOVA analyses the group means using F-distribution to assess if they are equal, and results in a significant result if they are not. Although, ANOVA cannot identify which groups are different.

Partial least squares discriminant analysis (PLSDA) is a supervised technique that identifies the variation between the groups of a dataset. PLSDA has an advantage over other methods, such as Mahalanobis distance approaches, as it can include a number of variables that exceeds the number of samples, therefore allowing better quantification and thus interpretation.[170] The objective of PLSDA is to model the response variables by a set of predictor variables,[171] for example, relating Raman spectra to their biochemical properties. PLS reduces the dimensions of a dataset by extracting latent variables, known as PLS components, from a dataset that contribute the most to variance between the response variables of different groups. Therefore, statistical tests to analyse the variance, such as analysis of variance (ANOVA), do not need to be employed to select components. PLS finds the direction in the spectral space that contributes to the most variance in classification. The data is transformed onto a new set of variables to allow emphasis of the similarities and the differences between groups in the data. PLSDA is a multivariate technique that analyses multiple variables simultaneously giving it an advantage over univariate techniques. Least squares regression is carried out on the component-reduced data to calculate the coefficients for prediction through applying a regression line that results in the least total sum of squares error. Discriminant analysis is supervised through assigning labels for the different groups in the model. The number of components used in the model is selected to avoid overfitting resulting from too many components, and avoid loss of discriminatory information that is caused from too few components. The optimum number of PLS components is chosen by testing the model, adding another component each time until the performance of the model starts to decrease, thus stopping at the point where the root mean squared error (RMSE) of prediction is at a minimum. Cross validation uses a blind dataset, unseen by the model, and compares it to the training dataset. It is performed to calculate the classification accuracy of the PLS model using a leave-one-sample-out approach in order to compare it to histopathological diagnosis. The process of PLSDA validation is shown in Figure 1.22 as a flow diagram.

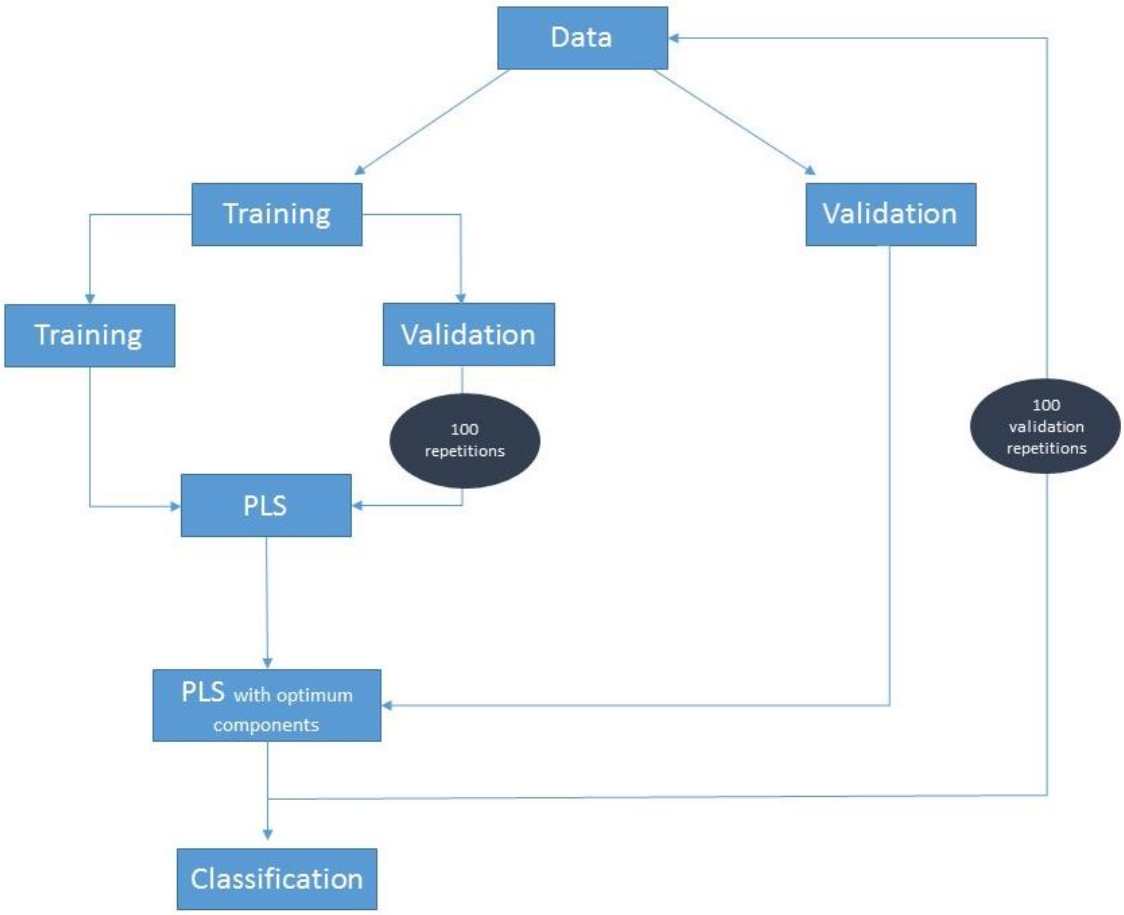

Figure 1.22- Flow diagram for validating PLSDA.

Variable importance in projection (VIP) highlights specific wavenumbers that are the most significant for discrimination and quantitatively defines their level of importance. Therefore, applying this technique can enable selection of specific wavenumbers in the spectrum, hence speed up data analysis and data collection, whilst also reducing file size for data storage. VIP applies a quantitative threshold in order to identify the most selected discriminatory variables. This enables impartial interpretation of the components to determine the molecular bonds in biochemicals that contribute the most to variance. Lloyd *et al.*[172] demonstrate the ability to generate similar classification sensitivities and specificities, using only the regions of the spectrum identified through VIP as being >95% important for discrimination, to using the whole spectral range.

The classification sensitivities and specificities were calculated from the multivariate models. Sensitivity represents the number of positive samples correctly classified out of the total number of samples for that pathology. Although

there is a continuum of disease progression for the stages of cancer, the type of cancer is categorical. The pathologist diagnoses the samples and this result is used throughout to categorise the data. Therefore, sensitivities and specificities are based on this gold standard, which is already known to be a subjective measure for determination. Only the follow up of outcomes compared to the histopathology would improve the reliability, and this is usually not possible. Also, the ease to classify disease is very dependent on the disease and anatomical location. For example, it is far easier for a pathologist to recognise metastatic disease than most primary cancers as a result of the presence of different cell types in a tissue specimen. Specificity represents the ability of the multivariate model, as a percentage, to correctly identify negative samples as non-cancerous.[173 174]

Sensitivity = [TP/(TP+FN)] x 100

Specificity = [TN/(TN+FP)] x100

Where:

TP= true positive (spectral number correctly classified as cancer)

FN= false negative (spectral number incorrectly classified as non-cancer)

TN= true negative (spectral number correctly classified as non-cancer)

FP= false positive (spectral number incorrectly classified as cancer)

1.5 Optical Diagnostic Studies

The diagnostic potential of optical spectroscopy has been investigated in many studies that employ tissue specimens of a variety of pathologies. A large number of studies have focused their investigations on the diagnosis of cancerous tissue. Numerous investigations into the diagnostic potential of Raman spectroscopy has been carried out on cancers of the: brain[87,175–178], gastrointestinal tract,[179–181] skin,[182–188] larynx,[189,190] breast,[9,191–194] cervix,[137,195] bladder,[196] prostate,[196,197] oesophagus,[85,126] Nasopharynx[198] lung[199] and colon.[85] However, reported studies on lymph nodes are limited. Lymph nodes are small in size and a high level of variability between nodes can result in their investigation being challenging.[200]

Smith *et al.*[201] conducted the first reported study on the assessment of the malignant spread of breast cancer to axillary lymph nodes. They observed that through Raman spectral mapping, the presence of molecules can be assessed across areas of lymph node sections. They identified the presence of carotenoids, proteins, nucleic acids and lipids, and suggest that these constituents can be compared across tissue features to further understand the biochemistry of breast carcinoma. However, they state the need for the acquisition of maps from a larger sample set, for more accurate comparisons to take place. Following this study, Smith *et al*.[202] investigated Raman spectroscopy for axillary lymph node classification. The study analysed 59 lymph nodes from 58 patients and reported 91% and 93% correctly classified positive and negative lymph nodes respectively. Spectral votes for each node generated a classification sensitivity and specificity of 93% and 100% respectively. They reported an absence of the 784 cm^{-1} peak, assigned to cytosine, thymine, uracil and O-P-O, in the spectra taken from negative nodes. Peaks at 1086 cm^{-1} and 1578 cm^{-1} were observed to be of a higher intensity and the 1329 cm^{-1} shoulder is more prominent in the spectra of positive lymph node spectra. They state that these spectral changes indicate a lower contribution from DNA in negative lymph nodes. A peak at 1184 cm^{-1}, assigned to tyrosine and possibly calcium fluoride and phenylalanine, is much more intense in the spectra of positive tissue. Peaks from collagen contributions are reported to be less intense in the spectra of positive nodes.

Horsnell *et al.*[75] investigated the *ex vivo* diagnostic ability of Raman spectroscopy for lymph node metastases in breast cancer with the use of radial based function-support vector machines (RBF-SVM). The study generated a classification sensitivity of 81% and a specificity of 97%, through the analysis of 10 spectra per node, from 16 positive and 31 negative lymph node specimens from different patients. This followed an earlier study by Horsnell *et al*.[203] which examined 38 axillary lymph nodes from 20 patients, of which 25 nodes were negative and 13 exhibited metastases. Through principal component-linear discriminant analysis (PC-LDA) Sensitivities and specificities were calculated as 85–92% and 88–100% respectively, analysing 5 spectra per sample and using a 'majority wins' approach to form a diagnosis. They note that the false positives produced in the study could be due to misdiagnosis that may occur as the result of unidentified micro-metastases, which are typically less than 12 mm in diameter, during histopathological diagnosis. They report that intra-operative assessment of lymph nodes would have an economic benefit of roughly £1368 per patient. However, spectral peaks with Raman shifts below 800 cm^{-1} were not included in the analysis. Similarly, Sattlecker *et al*.[204] also examined the diagnostic ability of Raman spectroscopy for the identification of breast metastases in lymph nodes. Through the analysis of 3385 spectra from 13 positive axillary nodes and 7092 spectra from 30 negative axillary nodes, a classification accuracy of 100% was produced using a radial based function-support vector machine (RBF-SVM) analysis. However, the lymph nodes being examined in this study were only from patients with breast cancer and other lymph node cancers were not investigated. They state that sufficient discrimination between cancerous and non-cancerous tissues could be achieved through the analysis of spectra with just 300 wavenumbers to the same testing performance as spectra with 1500 wavenumbers, reducing computing time. They observed that no specific spectral region has a higher impact on the model's performance than another spectral region.

Orr *et al*.[205] investigated the classification accuracy of Raman spectroscopy for a variety of head and neck lymph node diseases. Using a two-group cross-validation model they measured sensitivities and specificities, on a spectral level, of 90% and 88% for reactive lymph node spectra (n=1389) and 88% and 90% for cancerous lymph node spectra (n=1584) respectively. When

the spectra were averaged per node, sensitivities and specificities were calculated as 92% and 86% for reactive lymph node (n=53) classification and 86% and 92% for cancerous node (n=50) classification respectively. The classification accuracy of 4 group cross-validation model was assessed on a spectral level. Sensitivities and specificities were measured as 81% and 88% for reactive node spectra (n=1389), 93% and 99% for squamous cell carcinoma spectra (n=242), 76% and 95% for adenocarcinoma spectra (n=786), and 80% and 90% for lymphoma spectra (n=554) classification respectively. When spectra were averaged per node the sensitivities and specificities generally increased slightly. These were calculated as 85% and 86% for reactive nodes (n=53), 100% and 99% for squamous cell carcinoma nodes (n=8), 75% and 97% for adenocarcinoma nodes (n=24), and 83% and 92% for lymphoma node (n=18) classification. The data analysis they carried out consisted of normalisation and mean centring of spectra followed by multivariate analysis, employing principle component-linear discriminant function analysis. Although the lymphoma tissue was distinguished into either Hodgkin's or non-Hodgkin's pathologies, the ability of Raman spectroscopy to discriminate the two pathologies was not assessed in this study. Through a study of the diagnosis of benign, primary and secondary cancers in head and neck lymph nodes, Lloyd *et al.*[58] demonstrated the potential clinical efficacy of Raman spectroscopy. Through the *ex vivo* analysis of 123 lymph nodes from 23 patients, 2973 spectra were generated and analysed using partial least squares-discriminant analysis (PLSDA). Sensitivities and specificities were produced of up to 86% and 89% for benign neoplasm classification, 82% and 92% for secondary neoplasm classification, and 81% and 93% for primary neoplasm classification. However, whilst 18 patients had benign lymph node neoplasms, only 6 had secondary lymph node neoplasms and 3 had primary lymph node neoplasms.

Bird *et al*.[206] reported a "superior sensitivity" for cellular classification of breast micro-metastases in axillary lymph nodes, through the production of infrared micro-spectral images of 20 tissue specimens from different patients. They note that 7.33% of lymph nodes diagnosed as negative are later diagnosed as positive after re-examination; the detection of micro-metastases and isolated tumour cells (less than 0.2 mm in diameter) is difficult, yet vital for correct diagnostic staging of disease. Additionally, the ability of FTIR spectroscopy as a

diagnostic tool was demonstrated by Liu *et al.*[207] They reported a classification accuracy of 81.82% for papillary thyroid carcinoma micro-metastases in cervical lymph nodes, through the analysis of 184 lymph nodes from 22 patients, of which 14 were positive and 8 negative.

Isabelle *et al*.[208] measured 78,624 spectra from 9 benign lymph nodes and 11 cancerous lymph nodes to analyse the discriminatory power of both infrared and Raman spectroscopy. A training performance was reported of 98% and 94% for infrared and Raman spectroscopy respectively. They report an increase in nucleic acid concentrations in cancerous tissue when compared to benign tissue, in Raman spectra at peak 724 cm^{-1} attributed to CH_2 bonds and 1125 cm^{-1} attributed to C-C bonds. They state that the increase in nucleic acid concentrations support increased nuclear-to-cytoplasmic ratios and mitotic activity observed in neoplastic cells. A decrease in lipid and carbohydrate concentrations was observed in cancerous tissue compared to benign neoplasms. This was seen in Raman spectra at peak 1303 cm^{-1}, attributed to the CH bond and 1442 cm^{-1}, attributed to CH_2 bond of lipids, and peak 1084 cm^{-1}, attributed to the C-C bond in carbohydrates. It was also seen in the IR spectra at peak 1745 cm^{-1}, attributed to the C=O bond in lipids, and peak 1165 cm^{-1}, attributed to the C-O bond, and 1400 cm^{-1}, attributed to the COO^- bond in carbohydrates. They state that the decrease in lipids and carbohydrates support the use of these molecules as energy providers for proliferation in carcinogenesis.

Andrus and Strickland[209] reported that as the pathological grade of non-Hodgkin's lymphomas increased so did the intensity ratio of peak 1121 cm to 1020 cm-1 representing cellular RNA and DNA respectively. Their observations were made through the collection and analysis of FTIR spectra from nine snap-frozen sections of non-Hodgkin's lymphomas. They also found collagen peaks to increase, present between 1200 cm^{-1} -1700 cm^{-1}, relative to the 1084 cm^{-1} peak as the pathological grade of the lymphoma increased.

Chapter 2: Materials and Methods

2.1 Tissue Samples

Oesophageal, head and neck, and femoral lymph node samples were collected by surgeons at Gloucester hospitals NHS foundation trust, after ethical approval by the Gloucestershire local research ethics committee (REC references: 02/112G, 15/LO/1345 and 14/SW/1077 respectively).

Immediately after biopsy a portion of lymph node or an extra lymph node, not required as for diagnostic purposes, was placed in a cryovial and snap frozen using liquid nitrogen. The bulk lymph node sample was then stored in -80^{o}C freezers until tissue sectioning. The tissue samples were sectioned using a freezing microtome by an experienced histopathology technician. For the oesophageal and femoral samples two sections were cut from each bulk tissue block, one for spectroscopic analysis and another for pathologist diagnosis. The tissue section for spectroscopic analysis was cut at 8 µm onto calcium fluoride CaF_2 substrate (Crystran, UK). The CaF_2 substrates are circular windows with a 20 mm diameter and 1 mm thickness. The tissue section for pathological diagnosis was cut at 4 µm thick onto a glass microscope slide and stained with haematoxylin and eosin (H&E) in order to better show the tissue architecture and cells. For the head and neck lymph nodes three sections were cut for spectroscopic analysis. The first at 8 µm onto calcium fluoride CaF_2 for spectroscopic analysis. The second at 4 µm thick onto a glass microscope slide for histopathological review. And a third section cut at 15 µm for spectroscopic analysis to compare with the results from the needle probe measurements. The remaining node was used for needle probe measurements. The tissue sections on the CaF_2 slides were returned to the freezer and stored at -80^{o}C until required. Prior to measurement, the tissue sections were removed from the freezer and left to thaw for at least a half-hour before analysis. For FTIR spectroscopy, the sections were left for an hour to ensure any water had evaporated prior to measurements taking place.

2.1.1 Histopathological Analysis

Two pathologists from Gloucestershire Hospitals NHS Foundation Trust, Dr Suzanne Hopcroft and Dr Chandima de Cates, gave their expertise to this study to enable the pathological identification of the H&E stained tissue sections. The H&E stained tissue sections, adjacent to those spectrally measured, were microscopically examined in order for a diagnosis to be made. In cases of lymphoma it is difficult to make a definitive decision for frozen sections. In these cases the other half of the lymph node, taken by the surgeon and formalin fixed for a clinical decision, was paraffin set, sectioned and stained for analysis by the pathologist. Thus, the section looked at was not adjacent, however in cases of lymphoma, the cancer cells are not as localised as in metastatic and the lymphoma cells are expected to be throughout the tissue. Also, for samples where the pathologist was not completely sure, the section was taken to other pathologist to form a consensus.

2.1.2 Sample and Acquisition Numbers

The total number of samples collected for this study can be seen in Table 2.5. Twenty-two oesophageal lymph node samples were collected from fifteen patients: Fifteen nodes were identified as non-cancer (benign) from nine patients, and seven samples were identified to contain metastatic adenocarcinoma from six patients. Twenty-two femoral lymph node samples were collected from eight patients: sixteen were identified as non-cancer from five patients, and six were identified to contain squamous cell carcinoma metastases from four patients. A total of seventy-five head and neck lymph node samples were collected from sixty-three patients: forty-eight classified as non-cancer from thirty-nine patients, six as containing squamous cell carcinoma metastases from six patients, one as containing papillary carcinoma metastases from the thyroid, one as containing infiltrating adenocarcinoma from the breast, one as Hodgkin lymphoma, fifteen as non-Hodgkin lymphoma from fourteen patients and three reactive nodes from three patients caused by toxoplasmosis, cat-scratch and Castleman's disease. Some of the patients in this study had both non-cancer lymph nodes and metastatic ones.

Table 2.5- Total number of lymph node samples and patient numbers from the whole study.

	Head and Neck							Oesophageal		Femoral	
	Non-cancer	Metastatic			Lymphoma		Other	Non-cancer	Metastatic	Non-cancer	Metastatic
		SCC	PC	AC	HL	NHL			AC		SCC
Sample Number	48	6	1	1	1	15	3	15	7	16	6
Patient Number	39	6	1	1	1	14	3	9	6	5	4
Total Patients	63							15		8	

Not all the samples that were collected for this study were used for analysis. Reasons for this exclusion were: a patient infected by MRSA, poor measurements, nodes with too much adipose content leading to intense contributions of lipid peaks and the need for a more equal balance between the pathology groups. In the final situation, only one or two nodes would be used per patient to optimise patient numbers whilst trying to balance the pathology groups. A random selection of the nodes were chosen to represent the patient. Table 2.6 and Table 2.7 show the sample and patient numbers alongside the number of spectra used for the analysis of the 8 µm and 15 µm sections.

Table 2.6- Sample numbers and corresponding pathologies used for the 8 µm samples.

	Head and Neck		Oesophageal		Femoral	
	Non cancer	SCC	Non cancer	AC	Non cancer	SCC
Samples	13	5	6	6	8	6
Patients	13	5	6	5	5	4
Total Patients	18		11		6	
Point spectra	260	100	120	120	160	120
Map spectra	131738	44802	-	-	-	-

Table 2.7- Total number of lymph node samples used for the 15 µm study.

	Head and Neck						
	Non cancer	Metastatic			Lymphoma		Other
		SCC	PC	AC	HL	NHL	
Sample Number	39	6	1	1	1	14	3
Patient Number	38	6	1	1	1	13	3
Total Patients	62						
Map Spectra	53879	10330	714	1076	2690	30511	5151

2.2 Raman Spectroscopy

2.2.1 The Raman Spectrometer System

Raman measurements were taken using a Renishaw inVia confocal spectrometer system (Renishaw, UK) shown in Figure 2.23. The system employs an eyepiece with a magnification power of 10x/20 N.A. (numerical aperture) and also uses a camera in order to display the picture on the PC monitor. Spectra were generated using a long working distance NPlan objective (Leica, Germany) with a magnification power of 50x/0.5 N.A. A Leica NPlan objective with a magnification power of 5x/0.12 N.A. was used to locate areas of interest for measurements. Samples were excited using an 830 nm laser with an output power of around 300 mW at the source and approximately 150 mW at the sample. Laser wavelengths were chosen to minimise fluorescence and sample burning. A grating of 600 lines mm^{-1} was employed to collect spectra over wavelengths in the region of 460-2400 cm^{-1}. The system is equipped with a charge coupled device (CCD) detector to detect the scattered photons from the sample. The Raman system is controlled through a computer using the Renishaw software WIRE. A motorised XYZ stage provides ease for streamline mapping of samples.

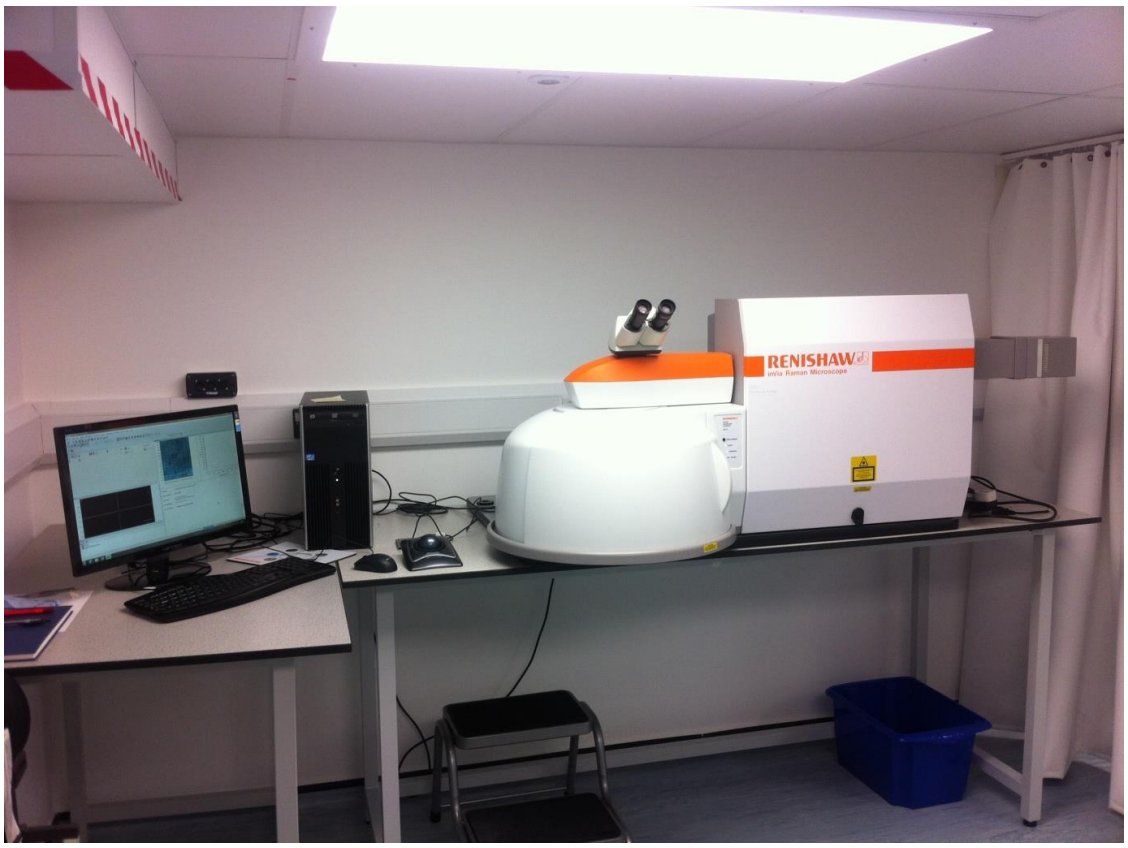

Figure 2.23-Renishaw in Via Raman spectrometer at the University of Exeter.

2.2.2 System Calibration and Standards Measurements

The system was offset against the 520 cm^{-1} peak of silicon (Si) prior to measurements taking place. Calibration is not ensured across the whole wavenumber range and therefore a Neon Argon lamp was measured also to evaluate the position of the emission bands. A measurement from green glass was also collected to correct for the energy sensitivity of the system. There is a known fluorescence for this glass and therefore the sensitivity of the CCD can be corrected using a transform function. An assignment table for biological signals frequently detected was collated and displayed in Table 2.8.

Table 2.8- Tentative assignment table for peaks observed in tissue spectra.

Tentative assignment		Peak position (cm^{-1})	Ref.
Vibration	**Biochemical**		
C-C twsit	Phenylalanine	622	6,153,190,210
C-C twsit	Tyrosine	645	6,190,210,211
C-S	Amino acid (cystine)	665	6,190,210,212
C-S	Cystine	694	213
C-N	DNA adenine/phospholipidhead	723	85,214
Ring breathing	DNA	748	215
Symmetric breathing	Tryptophan	755	6,85,190,204
Ring breathing	Pyrimidine base	784	6,58,85
	Phosphodiester (Z-marker)	809	212
PO_2^- stretch	Nucleic acids	827	6,85,215
Ring breathing	Tyrosine	851	155,214
C-C stretch	Collagen, hydroxyproline	874(w)	215
C-O-C	Adenine	899	6
C-C stretch	Protein backbone (α-helix conformation)	936	6,85,140
	Hydroxapatite, cholesterol	957	85,212
v(C-C) wagging	Collagen	970	212
Ring breathing	Phenylalanine	1001	6,85,140,190
C-H in-plane bending	Phenylalanine	1033	6,85,199
	Glycogen	1047	212,216
C-C stretch	Lipids	1060	85,212
C-N, C-C stretch	Protein, lipid, glycogen	1082	6,85,190,204,216,217
v(C-N)	DNA	1095	212
v(C-C), C-N	Lipid, proteins	1125	6,85,190,215
C-C, C-N stretch	Proteins/carotenoid	1154	6,155,190,214
C-H in-plane bend	Tyrosine	1172	6,190,212
v(C-C_6H_5)	Protein (tryptophan, phenylalanine)	1207	6,190,212
	Amide III	1261	6,85,215,218
CH_2 wag	Collagen	1282	219
CH_3/CH_2	Lipid/collagen	1302	85,199,212
CH deformation	Collagen	1320	214

CH deformation	Amide III	1338	214
CH	Tryptophan	1365	220
CH	Protein	1391(w)	221
CH_2 CH_3	Proteins, phospholipids	1445	6,190,212
CH	DNA	1490	217
r-C=C-r	Carotenoid	1527	85
C=C	Tryptophan	1551	204
Ring	Nucleic acids- Pyrimidine	1581	6,85
C=C bend	Amino acids (phe, tyr)/amide I	1604	6,155,190,218,222
C=C stretch	Amino acids (tyr, tryp)/amide I	1617	6,85,190,204,218,223
C=O/C=C stretch	Amide I: alpha helix (1650-57) Antiparallel beta sheet (1612-40) Parallel beta sheet (1626-40) Turn (1655-75, 1680-96) Unordered(1640-51)	1656	6,85,190,212,224
C=O Ester	Lipids	1733	214,225
C=O, C=C	Lipid	1747	85,204,226

2.2.3 Raman Measurements

Both point measurements and streamline maps were acquired for this project. Point measurements were acquired for 15 seconds at various locations after identification by the pathologist within the lymph node to ensure a representative area was sampled. Raman images were produced through streamline mapping of the sample with an acquisition time of 30 seconds per spectrum. In streamline mode, the time spent illuminating an area on the sample is the equivalent to the time divided by the fraction of the laser line that the mapping pixels approximate to. The streamline laser line used was 50 µm in length by 4 µm wide. Measurements of the 15 µm sections for pathological discrimination had a step size of 7.1 µm x 7.1 µm, and those at 8µm for comparison with FTIR images had a XY step size of 5.7 µm x 5.7 µm, in order to be more comparable to the FTIR step size of 5.5 µm x 5.5 µm. Therefore, the time the laser line illuminated the sample for are 4.2 seconds for the former and 3.4 seconds for the latter measurements.

2.2.4 Raman Needle Probe

The Raman needle probe that was employed in this study is described by Iping Petterson *et al.*[147] developed from an earlier probe by Day and Stone[146]. The probe used employs multiple optical fibres to enable maximum collection of scattered light. The optical fibres are composed of low-OH silica, with six collection fibres situated around one central delivery fibre. The fibres are 125 µm in diameter with a core of 105 µm and are all polished at the tip. The excitation wavelength employed in the system was 830 nm, delivered using a spectrum-stabilised laser module (FatBoy, Innovative Photonic Solutions, USA). A patch cable directs the excitation light through two collimating lenses surrounding two bandpass filters (BrightLine FF01-830/2-25, Semrock, USA) to the the needle tip. Maximum power at the sample was measured to be 230 mW, however the power was altered depending on the tissue type being measured. The scattered light is then collected and directed to collimating lenses (Thorlabs, F220FC-780 nm) to pass the light through two long-pass filters (RazorEdge, LP02830RU-25, Semrock, Rochester, NY, USA) and then to the spectrometer (RamanExplorer 830, 1002A-00465, Headwall Photonics, Fitchburg, MA, USA). A back illuminated, deep depletion charged-coupled device camera (iDus, DU420A-BR-DD, Andor, Belfast, UK) is coupled to the spectrometer for detection of the Raman scattering. The tip of the probe is a 22 gauge stainless steel tube 7 cm in length, to allow lymph nodes in the head and neck to be reached through the skin. The fibres are fixed in place in the tube using medically approved epoxy resin (Opti-tec 5007, INTERTRONICS, UK). A 19 gauge hypodermic needle acts as a sleeve around the tube to allow insertion of the needle prior to insertion of the fibre-containing stainless steel tube in order to minimise fibre damage and tissue debris building up before it reaches the target tissue. The filters and expensive optics are incorporated outside the needle probe tip to allow it to be disposable and thus sterile for each patient use. The probe tip is optically coupled to the handle using ceramic ferules (Alker Optical Equipment, UK). Figure 2.24 shows the schematic setup of the needle probe instrumentation design.

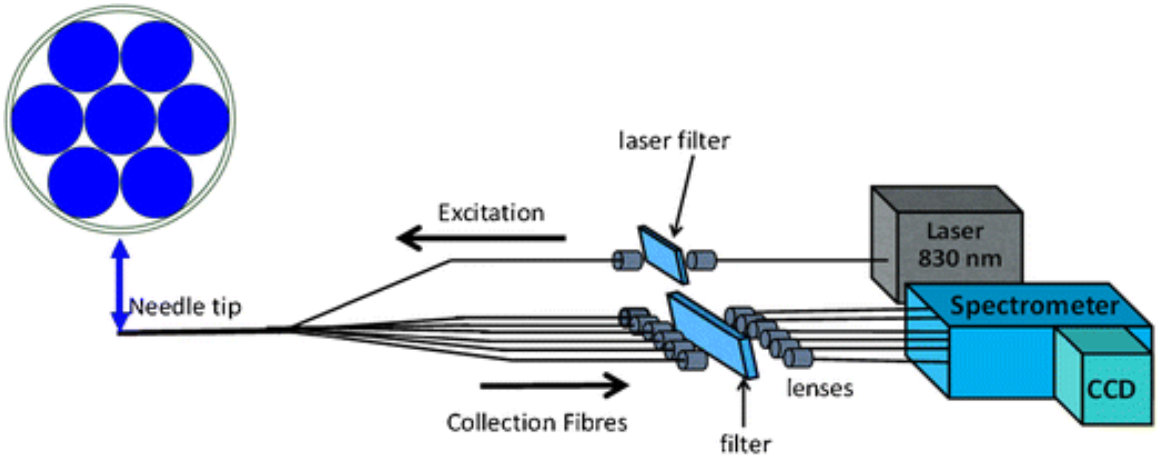

Figure 2.24- Schematic of needle probe instrumentation taken from reference.[147]

2.2.5 Tissue Simulating Phantoms

Intralipid and India ink phantoms were used as tissue simulating phantoms, due to their scattering and absorbing properties, to assess the performance of the needle probe. A phantom was prepared of 1% intralipid (20% emulsion, Sigma) and 0.005% India ink (Histology Stain, American MasterTech). The reduced scattering coefficient is calculated using the following equation:

$$\mu_s' = a(\lambda(nm)/500)^{-b}$$

From Jacques,[227] the 'a' and 'b' for soft tissue, such as lymph nodes and the surrounding tissue, is 18.9 and 1.286. With a laser wavelength of 830 the reduced scattering coefficient equals 9.8422 cm^{-1}. Vardaki *et al*.[228] show that 1 % intralipid is acceptable for a reduced scattering coefficient of 9.2 cm^{-1}.

Polytetrafluoroethylene (PTFE) was submerged in a bath of the intralipid and ink solution and a micrometre was utilised with the probe attached whilst enabling accurate alteration of the distance for every measurement. For the initial measurement the probe tip was in contact with the PTFE and then the distance between the two was increased by half a millimetre for each subsequent measurement.

Prior to acquisition of sample measurements, standards were taken. These included NeAr to calibrate the instrument, silicon for a reference peak,

stainless steel and green glass to later correct for instrumental contributions.

The remaining bulk head and neck nodes were measured using the Raman needle probe at 150mW of power at the sample. Once defrosted, the sample was placed on a CaF_2 window and the Raman needle probe was clamped in position above the node ready for measurement. The needle probe would then be lowered and 10 continuous 1 second measurements would be taken. This was then repeated at various locations on the sample so at least 10 areas were measured.

2.3 Fourier Transform Infrared (FTIR) Spectroscopy

2.3.1 The FTIR Spectrometer System

Infrared measurements were taken using an Agilent Cary 600 series Fourier transform-infrared imaging (FTIR) spectrometer system, coupled with an FTIR microscope (Agilent Technologies, USA), and shown in Figure 2.25. The microscope system consists of a wide field eyepiece magnification power of 10x/0.18 N.A and measurements were taken using an Agilent objective with a magnification power of 15x/0.62 N.A. A 128 x 128 liquid nitrogen cooled focal plane array (FPA) detector was employed in the instrument, which allows simultaneous detection of spectra. Transmittance mode was utilised for this study and an under-sampling ratio (UDR) of 4 was included to filter the stray light.

Figure 2.25- Agilent Technologies Cary 600 series Fourier Transform Infrared Spectrometer and microscope system at the University of Exeter.

2.3.2 FTIR Measurements

The spectral resolution of 4 cm^{-1} was selected for the FTIR measurements. 256 scans were summed for the background spectrum and 64 scans were co-added for each tissue spectra. FTIR images were acquired to measure a large proportion of the tissue sections. The tissue sections used were those cut at 8 µm rather than 15 µm, so the light could be transmitted through the sample more readily. The pixels in the image are 5.5 µm x 5.5 µm. The FPA camera pixels are 40 x 40 $µm^2$ in size, internal microscope objectives change the magnification of the sample that is passed onto the camera and results in the image magnification of approximately seven times in standard magnification mode with the 15 x objective.

2.4 Data Analysis

Analyses of the data were carried out using the computer software Matlab R2015b (The MathWorks, Inc, USA). The analytical methods for the data was chosen to both align with literature[58] and also as a result of tested algorithms with research group.

2.4.1 Pre-processing

The signal-to-noise ratio (SNR) is used to characterise the noise of the spectrometer system being used and is especially important for *in vivo* application where the integration time is limited for clinical practice. The following equation was used to calculate the signal-to-noise in order to assess the quality of the signal collected.

$$SNR = \overline{S}/\sigma_y \qquad \text{Eq. 2.4.1}$$

Where:

$\sigma_y = \text{Standard deviation of the peak height}$, $\overline{S} = \text{Average peak height of a peak}$

Spectral wavelengths were cut down to only include the fingerprint region over 600-1800 cm^{-1}, to speed up analysis but retain a reasonable amount of spectral information. Pre-processing was carried out on the data initially to correct for unwanted spectral features. Raman data were green glass corrected, and vector normalised. Median filtering was also applied on map data to remove unwanted cosmic rays, where each acquisition is not visually checked for these contributions. Green glass acquisitions are taken after calibration and prior to every image, or set of sample measurements, to enable correction of the instrument response function (which can change subtly when the system self-aligns, moving the edge filter set). For Raman maps, median filtering was applied across three pixels. The median was calculated where the two adjacent pixels, to the pixel of interest, are used to replace the pixel to remove the unwanted cosmic

ray whilst not losing that pixel from analysis. The data were finally vector normalised to account for intensity differences caused by variations during data collection. For visualisation of the spectra, baseline subtraction and smoothing was applied to the data, but this method was not used for multivariate analysis. A third order polynomial fit was applied for baseline subtraction in this study as it was observed to be the optimum for reducing the baseline without altering the spectral features. The Savitzky-Golay algorithm was employed to smooth the data, using a 3rd order polynomial fitted across 7 wavenumbers, to reduce the noise contributions. A region of interest (ROI) was selected from the spectral maps based on pathological review of the adjacent H&E section.

For the FTIR data the only pre-processing that was carried out was vector normalisation and mean centring prior to multivariate analysis.

2.4.2 Multivariate Analysis

Once pre-processed, the data were mean-centred before any further analysis. Principal component analysis (PCA) and for maps K-means cluster analysis was applied to observe unsupervised spectral grouping and assess the peaks in the spectra that contribute to variation of the dataset. Multivariate analysis was then carried out to create and evaluate classification models for identifying pathological groupings of the dataset. Partial least squares-discriminant analysis (PLSDA) was employed to create the classification model and a leave-one-sample-out cross-validation was carried out to assess the accuracy of the model. The components selected for the models were between 3 and 7, depending on when performance of the model was at an optimum. A majority wins approach was used, similar to how it would be classified in a clinical setting, to identify groups depending on all the measurements from one sample. Figure 2.26 shows the flow diagram for the data analysis of the spectral measurements.

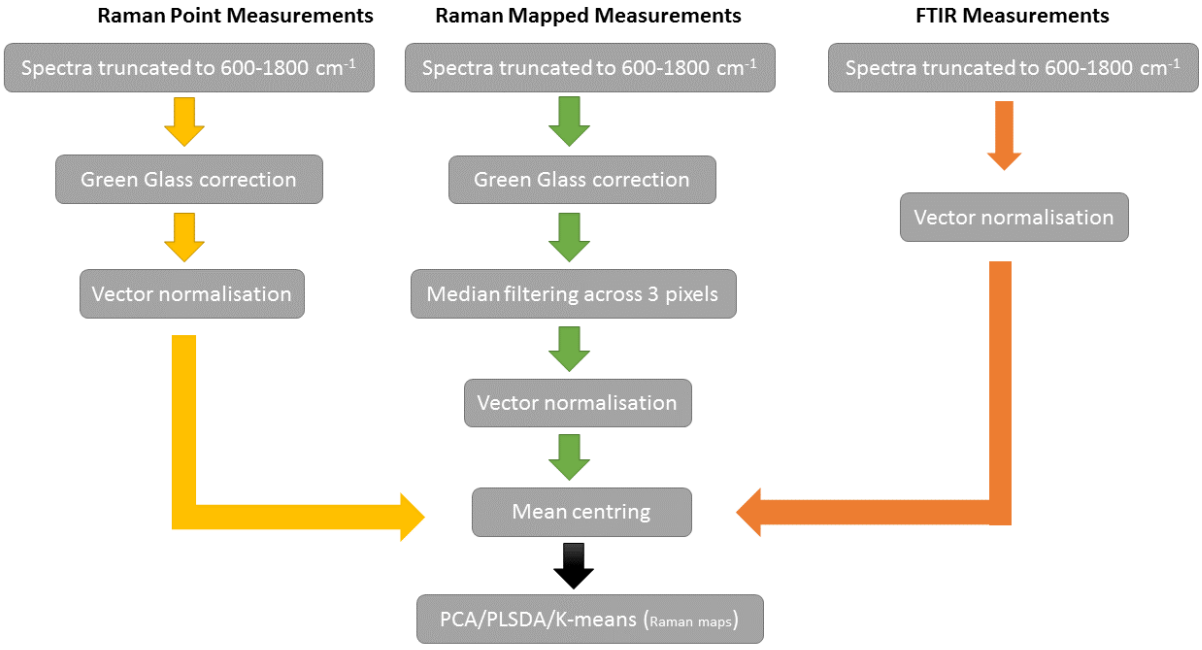

Figure 2.26- Flow diagram of the order of pre-processing and methods of data analysis.

2.4.3 Data Analysis of Needle Probe Measurements.

The green glass measurements provided a calibrated fluorescence spectrum to correct for instrumental response. The green glass could also be used to correct for etaloning of the CCD. The data were then cropped to 575-1800 wavenumbers and interpolation was carried out. Any spectra that were above the overall mean plus 1 standard deviation was removed from analysis in order to identify and eliminate saturated spectra. The data were then normalised and mean centred. Orthogonal partial least squares (OPLS) was carried out to identify components in the spectrum that cause variation, yet are not as a result of biochemical changes. Four components were identified, these included spectral features present as a result of fibre signal, etaloning, background and intense lipid peaks. Hoteling T-squared (a t-test in multiple directions) was performed to remove outliers outside an ellipse calculated from the scores at 95% (2 standard deviations). PLSDA was then carried out on OPLS dataset and a leave-one-out cross-validation was performed to predict classification accuracy.

For visualisation, the data were baseline corrected and the etaloning was corrected using a 2nd order polynomial fit of the stainless steel measurement using the etaloning shape from the green glass.

Chapter 3. Raman Spectroscopy of Metastatic and Non-Cancer Lymph Nodes from Different Anatomical Locations

The lymph nodes collected and measured for this study were from the oesophageal, head and neck and femoral regions and consisted of a variety of pathologies. Lymph nodes are frequently some of the first organs metastatic cancer cells will travel to and therefore it is of interest to assess the spectroscopic differences between the metastatic and non-cancer lymph nodes in this study, to evaluate the clinical abilities of Raman spectroscopy for this application. In this chapter, the identification of secondary cancers from benign tissue will be assessed, both as lymph nodes from their individual regions and also as lymph nodes combined.

3.1 Results

3.1.1 Raman Spectroscopy of Oesophageal Lymph Nodes

During experimentation it was observed that oesophageal lymph node measurements possessed much higher lipid signal than many other biological tissue samples. Measurements were acquired of the surrounding fat tissue in order to obtain its spectral signature and clearly observe the lipid peaks that contribute. Figure 3.27 shows the spectrum of fat averaged across four measurements for more accurate representation of the signal. Lipid peaks are seen at 1080 cm^{-1} with a shoulder at 1065 cm^{-1}, 1119 cm^{-1}, 1300 cm^{-1}, 1439 cm^{-1} and 1745 cm^{-1}, attributed to C-C, CH_2, C=C and C=O molecules within the tissue.

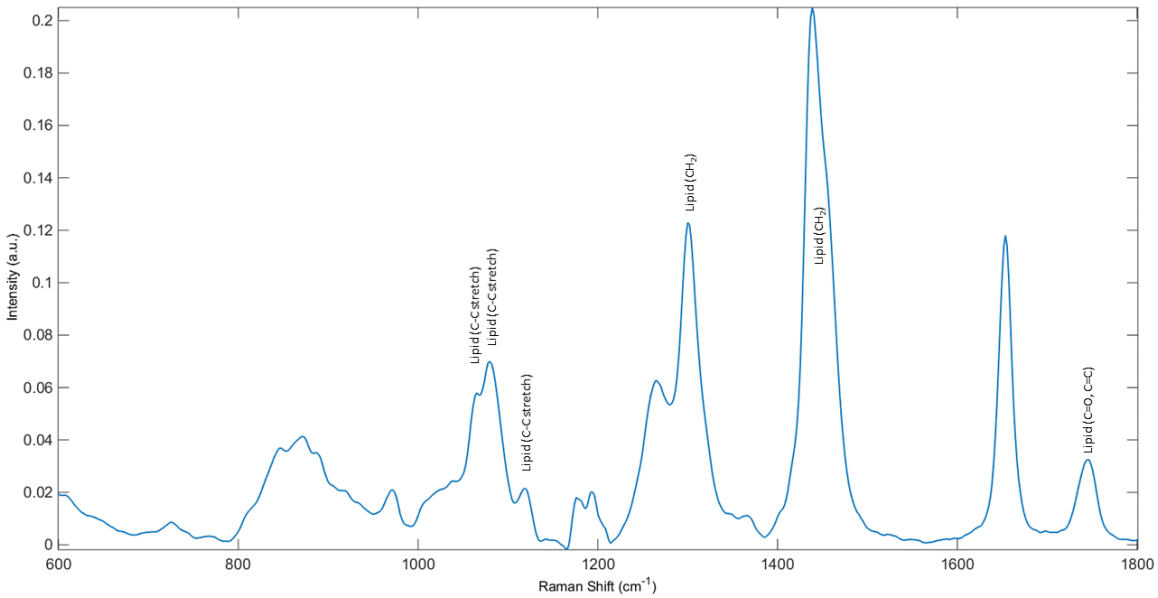

Figure 3.27- A spectrum from the adipose tissue surrounding the lymph nodes averaged across four acquisitions with lipid peaks labelled.

Dark coloured microscopic particles were seen in most of the oesophageal lymph node tissue sections, shown in the images of an H&E stained section and an unstained section in Figure 3.28. These particles are carbon pigments which pathologists diagnose as anthracosis.

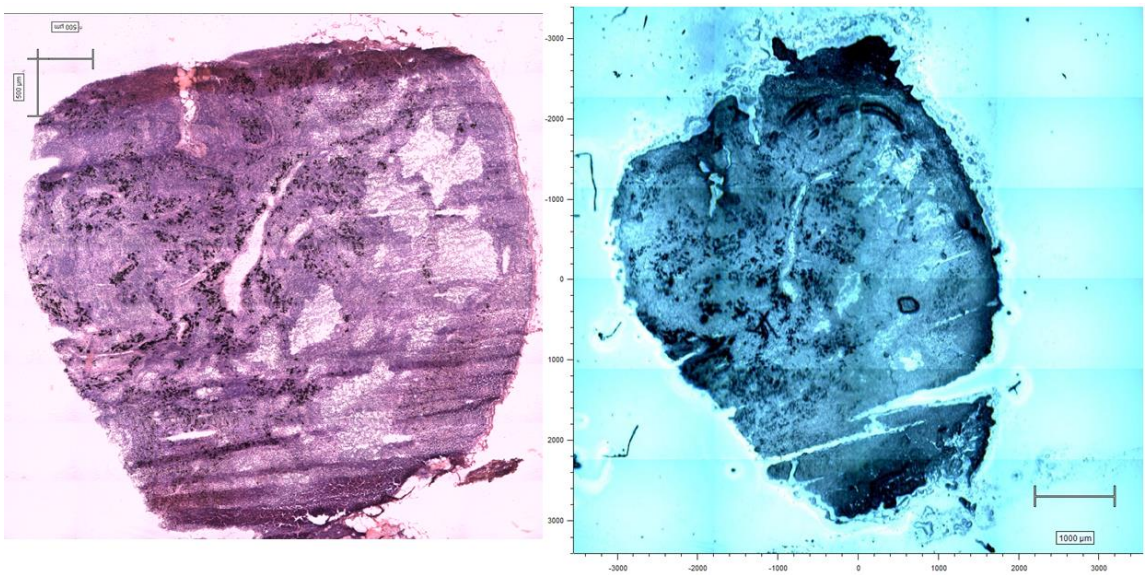

Figure 3.28- H&E (left) stained section of a lymph node section and a white light (right) image of the subsequent section from the same lymph node exhibiting anthracosis in the form of black pigmentation.

These heavily pigmented areas in the section would often result in a very strong spectral background when acquired for the time needed to get a biological spectrum with an acceptable signal to noise ratio. Hence the acquisition of spectra was more difficult, especially when streamline mapping, where measurements are acquired whilst the stage is moving in order to reduce the time taken to map a sample. Thus, one spectral measurement is a combination of scattered light from all the tissue illuminated within the set acquisition time. If a fluorescing particle is illuminated it will cause an intense spectral background of that measurement, often masking the biological peaks of interest. Figure 3.29 shows a spectrum from the carbon pigments measured for just 1 second in order not to flood the detector. Due to poor signal-to-noise, as a result of the limited acquisition time, the spectrum was smoothed using a Savitsky-Golay filter over 10 wavenumbers. The graphite peak (G band) and disordered carbon peak (D band) are displayed at 1583 cm^{-1} and 1300 cm^{-1} respectively, attributed to the C-C bonds that make up the molecule.

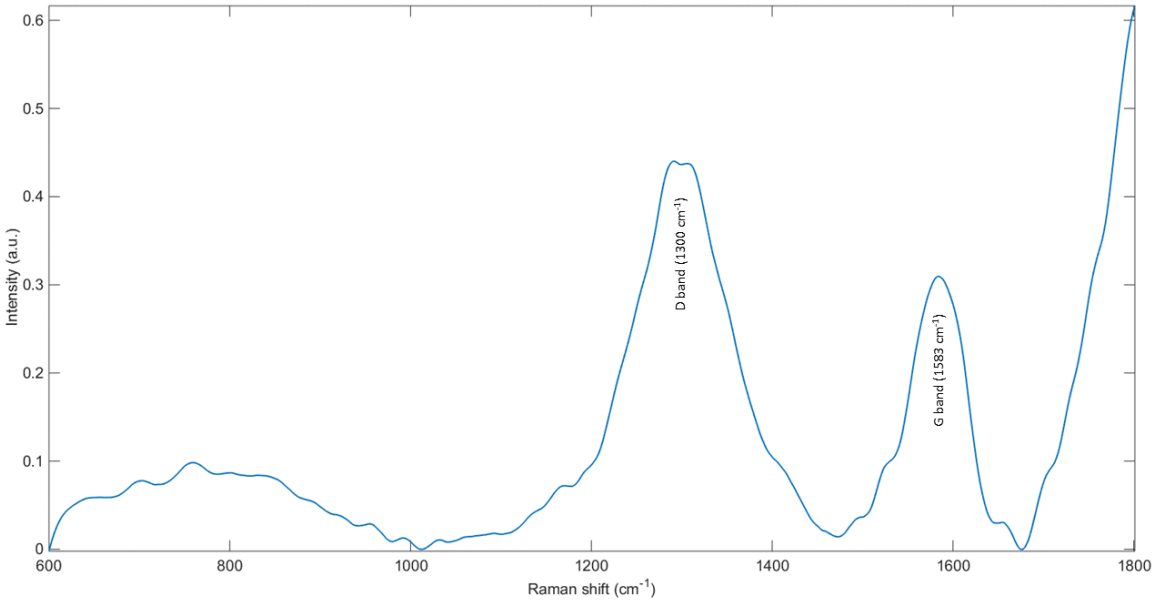

Figure 3.29- Carbon spectrum in a lymph node acquired for 1 second exhibits the D and G bands at ~1300 cm-1 and ~1583 cm-1.

Frozen samples can often result in scoring marks across the tissue section from the blade when they have been cut using a cryo-microtome. These scoring patterns can often make direct comparison difficult between the H&E stained sections and their subsequent spectroscopic section. An example of typical scoring patterns often seen on a tissue section is shown in Figure 3.30. Samples that have been formalin fixed and paraffin embedded prior to sectioning, are much less damaged and striated by the sectioning process. However, for spectroscopic analysis it is preferred for tissue samples to be in a state similar to how they would be *in vivo*. Therefore, the samples are put in liquid nitrogen in order to snap-freeze them, to avoid crystallisation of the tissue, and preserve them through low temperatures rather than external chemicals, even at the expense of striated sections.

Figure 3.30- H&E section of lymph node tissue exhibiting diagonal scoring patterns caused by the blade of the microtome.

To assess the spectral heterogeneity of lymph node samples, a Raman map was acquired in order to produce a spectral image. A section of lymph node excised from underneath the diaphragm, where carbon pigmentation is less common, and possessing identifiable nodal regions was measured using Raman streamline mapping. Figure 3.31 shows the microscopic image of the H&E stained tissue adjacent to the section mapped. The measured area is highlighted by the black box on the image and the various nodal regions are labelled.

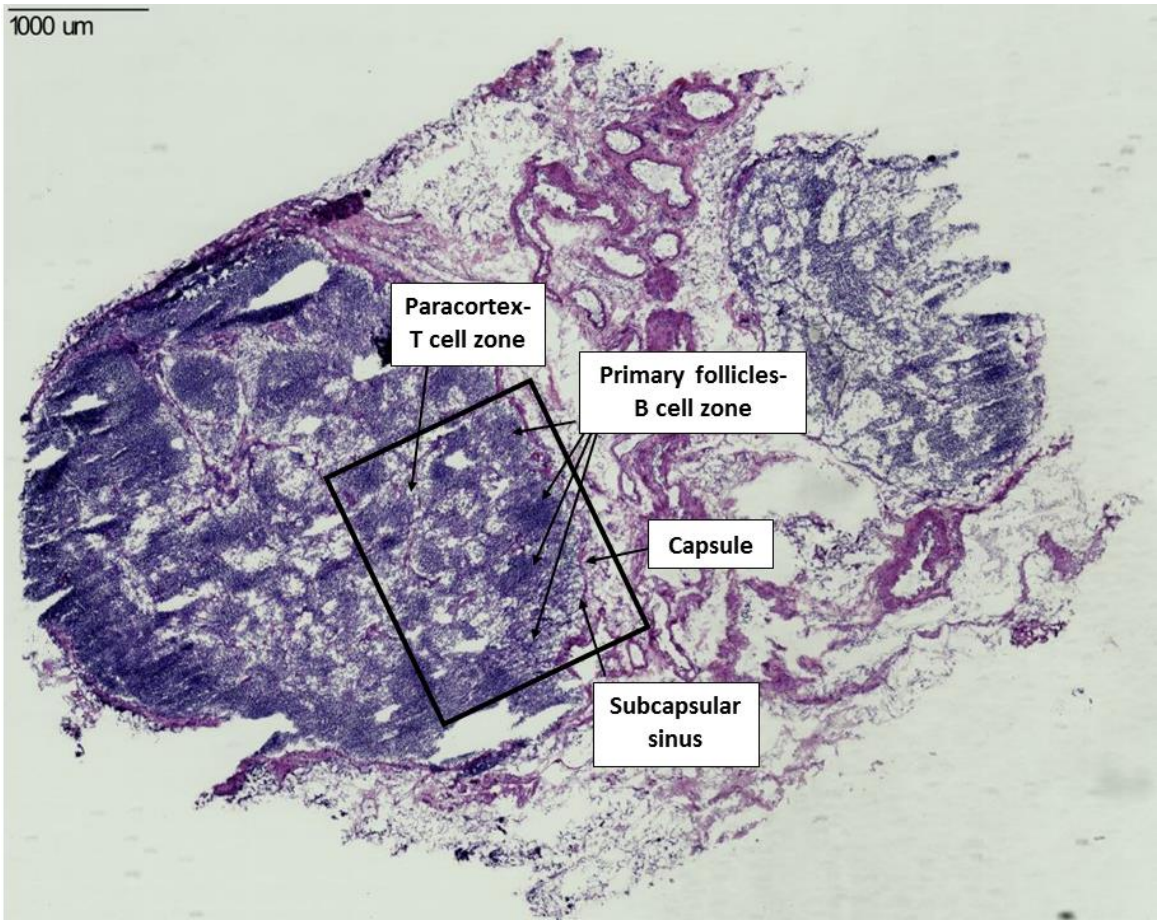

Figure 3.31- H&E stained tissue section showing the different regions within the oesophageal lymph node. The region in the black square represents the area mapped using Raman and FTIR spectroscopy.

Raman spectroscopy was utilised to collect a spectral map to produce a pseudo-coloured image, shown in Figure 3.32, through 6-cluster k-means analysis, an unsupervised technique where clusters are formed based on spectral similarities. Each pixel on the image represents a different spectrum. Using the H&E section, the colours from the k-means image can be assigned to

the various regions of the lymph node. It would appear that the follicles containing B-cells are represented in yellow, the surrounding T-cells in turquoise, the capsule in light blue, missing tissue in orange, lipid in maroon and dark blue represents the subcapsular sinus and other areas that are not as densely populated with lymphocytes. In order to see the spectral differences, the averaged spectra of the T-cell and B-cell regions are displayed in Figure 3.33.

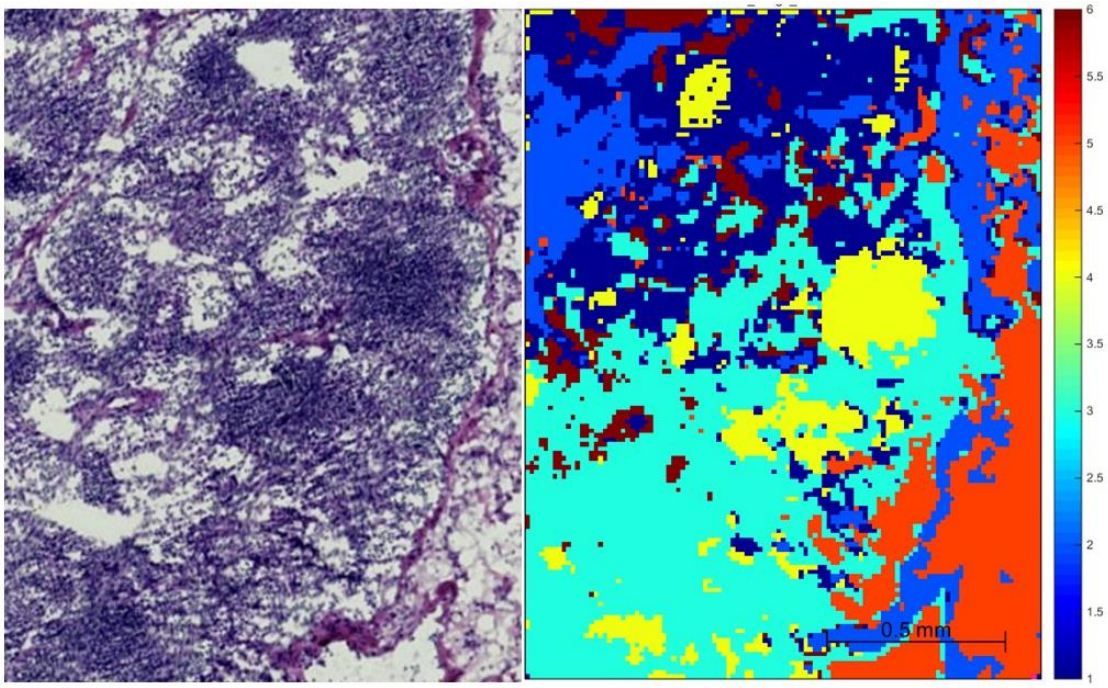

Figure 3.32- Pseudo-coloured image of 6 cluster K-means analysis on a Raman mapped non-cancerous lymph node section (right) with the corresponding region of the H&E stained tissue.

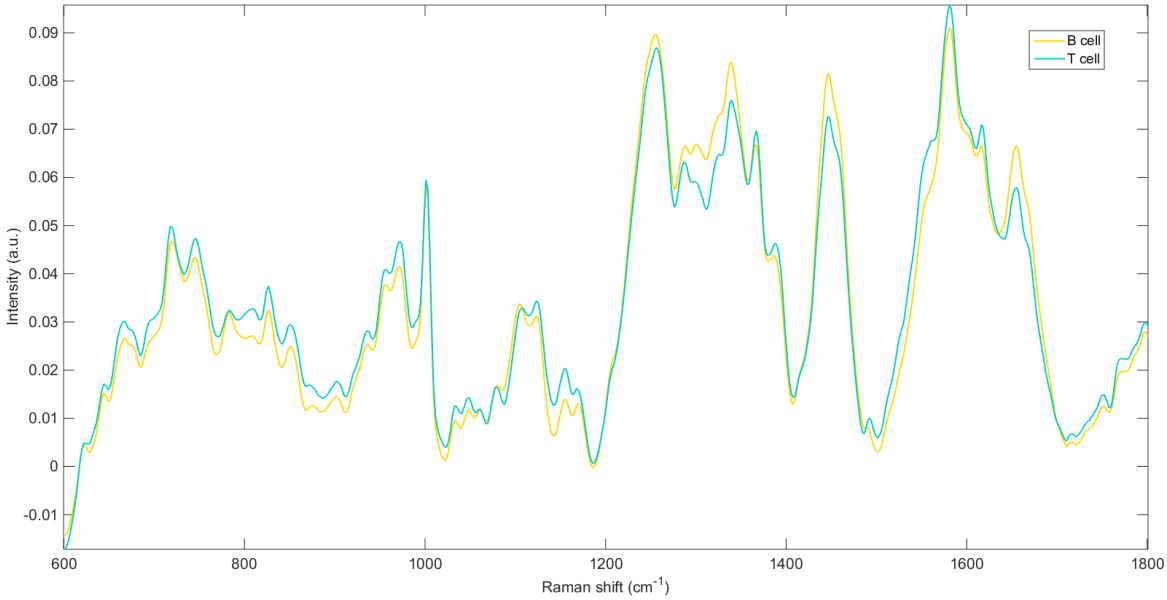

Figure 3.33- Mean spectra of the T cell area (blue) and the B cell area within the follicles (yellow) from the correlating colour on the k-means cluster image in figure 6.

The oesophageal lymph node samples only consisted of non-cancerous tissue and metastatic adenocarcinoma originating from primary cancers in the oesophagus. Therefore, the spectra from only two cell types were analysed: non-cancerous lymphocytes and cancerous epithelial cells of glandular origin. A pathologist highlighted any metastatic cells and the nodal regions on the adjacent H&E stained section. This enabled spectral acquisition of the specific metastatic areas when measuring from the cancerous lymph nodes. When collecting point spectra from the non-cancerous sections, spectral measurements from different areas within the lymph node were taken to have a representative dataset of the whole lymph node.

Due to heavy lipid and carbon contributions in the samples, spectral maps could not be acquired from many samples without their spectral contamination. Therefore, point spectra were taken in order to avoid areas that did not yield the appropriate biochemical information in their spectra. Samples that were heavily contaminated by lipid or carbon, that affected the resulting spectra, were left out of analysis. Also, there is a large discrepancy between the sample sizes of groups. To avoid a group being too over-represented, samples were sometimes

omitted from groups if there were multiple biopsies from the same patient, to only include one lymph node per patient.

To aid visualisation of the spectral differences between measurements from non-cancerous lymph nodes and those with adenocarcinoma (AC) metastases, the mean spectrum was calculated for each pathology and displayed in Figure 3.34. In typical spectra from biological tissue, the main peak contributions are a result of molecular bonds within amino acid, protein, lipid, nucleic acids and carbohydrate groups.

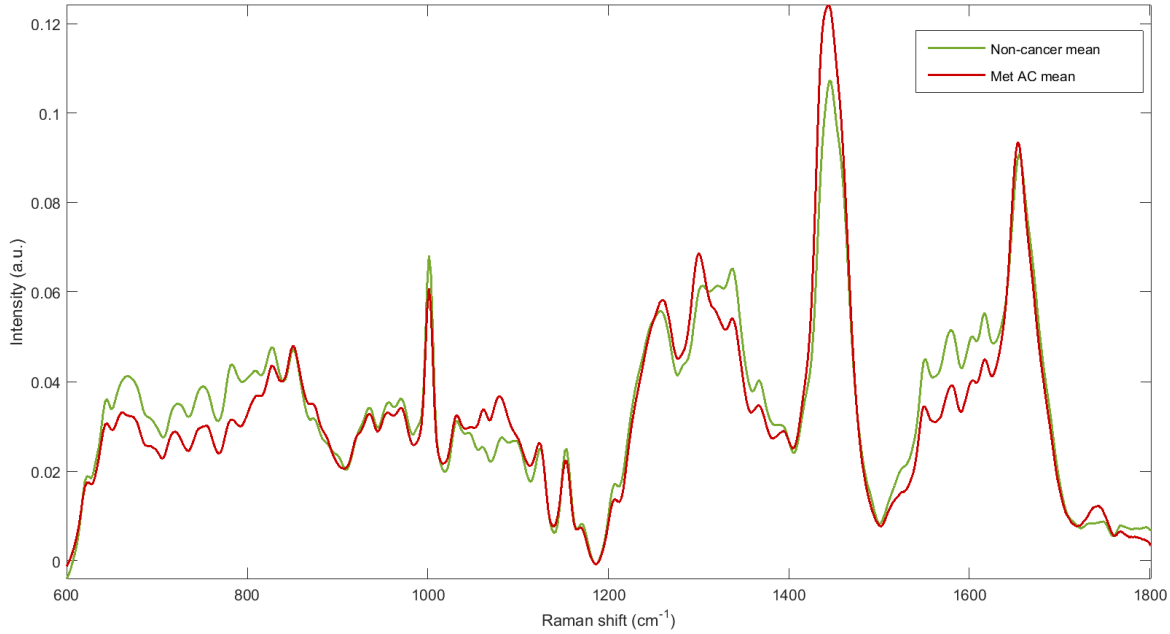

Figure 3.34- Spectra averaged across 140 AC met point measurements (red) and 280 non-cancerous point measurements (green) from 7 metastatic nodes from 6 patients and 14 non-cancerous nodes from 9 patients. The spectra are pre-processed and vector normalised and the band assignments are labelled.

Visual spectral differences between the two pathologies are primarily due to a variation in peak intensities. The protein, nucleic acid and amino acid peaks, found at wavenumbers 600 cm^{-1} to 810 cm^{-1}, 1336 cm^{-1}, 1365 cm^{-1}, 1513-1617 cm^{-1}, are more intense in the spectra from non-cancerous tissue than metastatic AC tissue. Lipid peaks are conversely more intense in metastatic AC than non-cancerous tissue, seen at 1061 cm^{-1}, 1078 cm^{-1}, 1298 cm^{-1}, 1439 cm^{-1} and 1743 cm^{-1}. The amide III peak at 1261 cm^{-1} and the amide I peak at 1654 cm^{-1} are also

more intense in the metastatic tissue measurements than in the non-cancerous sample measurements along with the collagen peak at 872 cm^{-1}. A carotenoid peak at 1523 cm-1 is seen to be more intense in the spectra of healthy tissue than that of metastatic tissue.

To show the variation between measurements from the same pathological group, the standard deviation is plotted around the mean in Figure 3.35. Although there is a large amount of variation between the dataset, the DNA peaks seem to be consistently more intense in non-cancer tissue measurements whilst those assigned as lipid peaks are always much more intense in metastatic tissue measurements. The signal-to-noise ratio was most commonly between 11 and 13 and greater than 5.71 in all measured cases. Therefore, most variation is attributed to biochemical differences between the samples rather than noise. The spread of data appears to be greater in the metastatic LN measurements than the non-cancer LN measurements. The dataset is balanced between the two groups so this may suggest a larger amount of biochemical heterogeneity between metastatic samples.

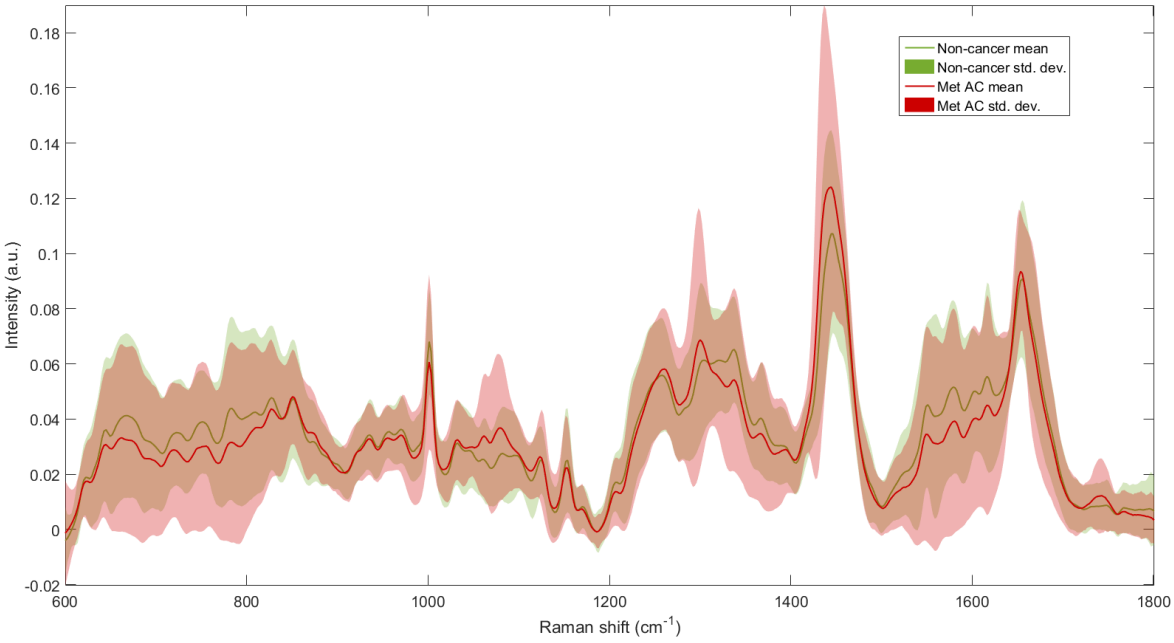

Figure 3.35- Averaged AC met (red) and non-cancerous (green) spectra. The shaded area flanking the spectra represents 2 standard deviations from the mean. Note that when both green and red overlap they are shown in brown.

The variations between spectra from metastatic AC tissue and non-cancer tissue are visually subtle and multivariate analysis was needed to identify spectral differences between the two pathological groups. Biochemical differences exist between pathological groups, therefore a supervised technique was employed to help identify the differences between the groups as oppose to the dataset as a whole. PLSDA was carried out to identify the components in the spectra that lead to the most variation between groups. The scores plot is displayed in Figure 3.36 based on the first three components. Analysis focused on where the scores were plotted, as either on the positive axis, above 0, or on the negative axis, below 0. In the first PLS component, the metastatic scores are generally negative whilst those from non-cancer measurements are more often positive. For PLS component 2 the non-cancer scores are more frequently positive and the metastatic, whilst more heavily negative, have a much broader distribution across the component. For PLS component 3 the metastatic scores are primarily negative and the non-cancer scores positive.

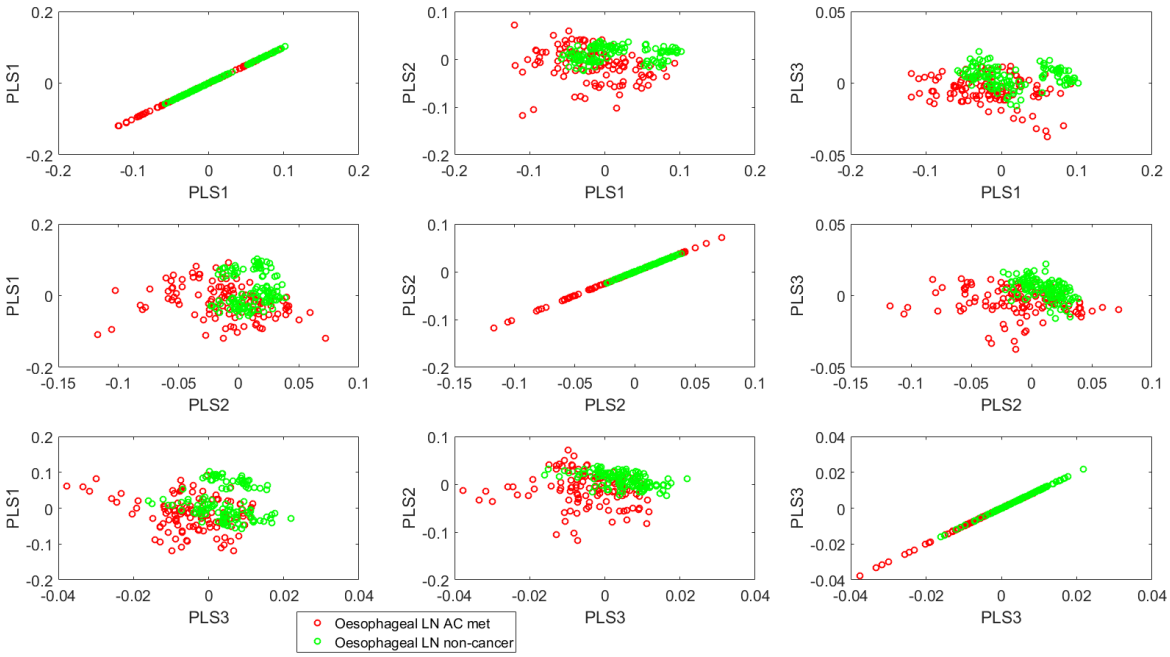

Figure 3.36- PLS score plot showing the grouping of AC met lymph node measurements (120 spectra), in red, and non-cancerous lymph node measurements, (120 spectra) in green, based on PLS components 1, 2 and 3.

Using a ‘leave one sample out’ cross-validation, the pathological group of the blind dataset were estimated. Classification sensitivities and specificities were

calculated between 50-83% for identifying metastatic lymph node data correctly from non-cancer measurements using a majority wins approach per sample. The optimal number of PLS components was calculated to be primarily 5 for this cross-validation. Table 3.9 shows the confusion matrix for this cross validation which uses a majority wins approach for measurements per node.

Table 3.9- Table displaying the confusion matrix and calculated sensitivity and specificity from blind classification using the PLS model from figure 10.

Confusion matrix		**True group**	
		Metastatic AC	**Non-cancer**
Predicted group	**Metastatic AC**	3	1
	Non-cancer	3	5
Total LN Samples		6	6
Sensitivity		0.5000	0.8333
Specificity		0.8333	0.5000

The PLS weights from the components were analysed in order to gain more information regarding the biochemicals responsible for the variation between the metastatic and benign tissue. Figure 3.37 shows the spectra of these weights and Table 3.10 lists the main peaks as either a positive or negative band. These weights can be compared to the scores on the PLSDA plot which represent the component weights based on the level of peak contribution in the spectrum. Therefore the direction of the peak, either positive or negative, can be compared to the PLSDA scores of the data, in order to identify which pathological group it is more characteristic of.

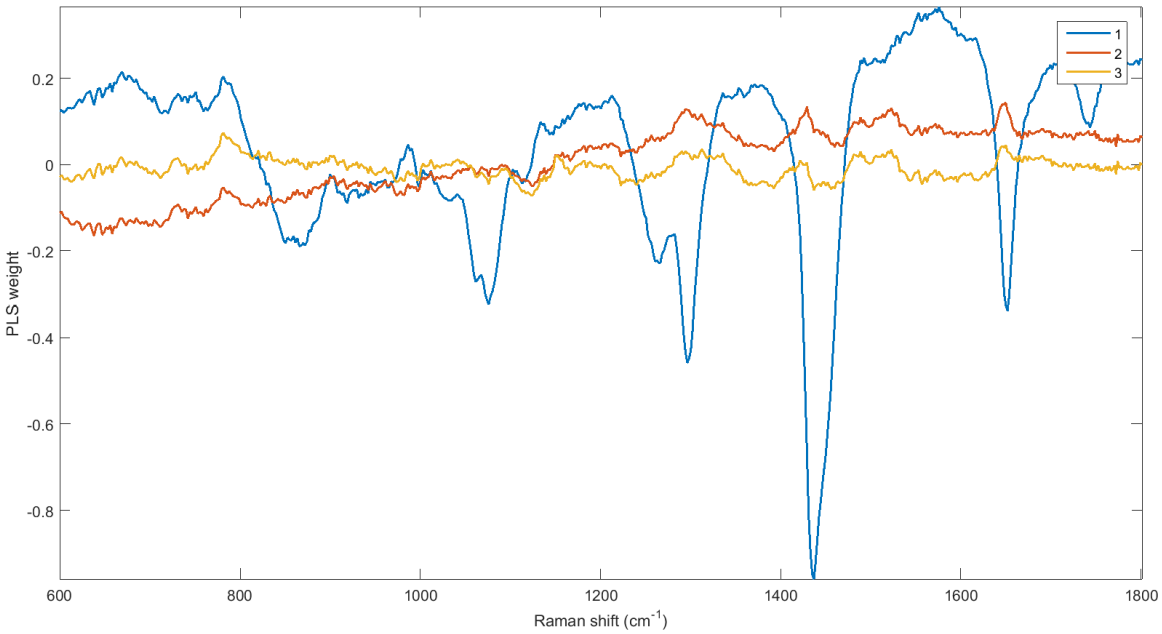

Figure 3.37- Spectra representing the weights of PLS components 1 (blue), 2 (purple) and 3 (green).

The weights from PLS component one support the observation that there are more intense lipid peaks in metastatic AC LN tissue, seen by peaks in the negative direction at 1062, 1076, 1296, 1437 and 1743 cm^{-1}, resulting from the C-C, CH_2CH_3, C=O and C=C bonds. Amide III is assigned to the 1263 cm^{-1} peak, present as a negative band in component 1 along with a peak at 1652 cm^{-1} assigned as amide I, supporting the observation of higher levels in metastatic AC tissue. Peaks assigned to DNA and amino acids are observed in the positive direction, such as at Raman shifts of 669, 731, 750, 781, 1553 and 1557 cm^{-1} caused by vibrations of C-N bonds and aromatic ring structure, which are components of DNA molecules and nucleic acids. This is also in agreement with the observations that the DNA seems to be present in lower quantities in the metastatic AC LN tissue than compared to the non-cancer LN samples. Table 3.10 lists the main peaks from the component weights with their tentative biochemical assignments.

Table 3.10- Table of positive and negative bands from the weights of PLS components 1, 2 and 3 in figure 11.

Component 1				**Component 2**				**Component 3**			
-ve band		**+ve band**		**-ve band**		**+ve band**		**-ve band**		**+ve band**	
851	Tyr	669	Cystine	974	Collgen	731	DNA	1123	Lipid/protein	731	DNA
865	Collagen, hydroxyproline	731	DNA	982	Collagen	781	Pyrimidine base	1437	Protein/ phospholipid	781	Pyrimidine base
1062	Lipid	750	Tryp			1152	Protein/carotenoid	1456	Protein/ phospholipid	1152	Protein/carotenoid
1076	Protein, lipid	781	Pyrimidine base			1294	Cytosine[229]			1294	Cytosine[229]
1263	Carb/amide III	1133	Protein, lipid			1429	Protein/ phospholipid			1429	Protein/ phospholipid
1296	Lipid/ collagen	1212	Protein			1489	DNA			1489	DNA
1437	Protein/ phospholipid	1553	Tryp			1523	Carotenoid			1523	Carotenoid
1652	Amide I	1576	Nucleic acid			1650	Amide I			1650	Amide I
1743	Lipid										

In the weights of PLS component two, negative bands at 974 cm^{-1} and 982 cm^{-1} are assigned to collagen due to vibrations of C-C and CH_2 bonds within the molecule, implying higher collagen in the LN with metastatic AC. The band at 1523 cm^{-1}, caused by r-C=C-r bond, assigned to carotenoid molecules is also present in PLS component two in the positive direction belonging mainly to the non-cancerous lymphocyte group, which is also seen to be different between the two pathologies in the mean spectra. In the positive direction a peak present at 781 cm^{-1} is attributed to the ring structure of the pyrimidine bases, further supporting their stronger presence in non-cancerous tissue. In the weights of PLS component three, lipid peaks seen at 1123, 1437 and 1456 cm^{-1} are negative and peaks at 781 and 1489 cm^{-1} are positive, and indicative of the vibrations of DNA molecules. Carotenoid assigned peaks at 1152 and 1523 cm^{-1} are also positive, along with an amide I assignment at 1650 cm^{-1}.

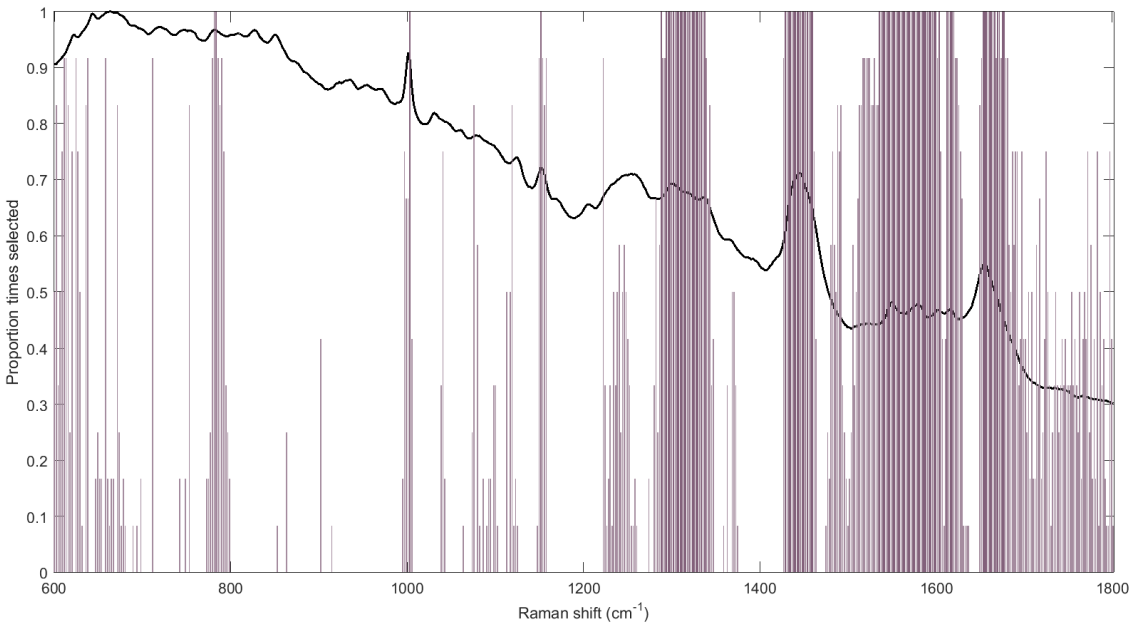

Figure 3.38- Bar plot showing the selection frequency of bands using VIP overlaid on the mean spectrum of both metastatic and non-cancerous data.

VIP was used to identify the important Raman shifts for discriminating between the groups. Figure 3.38 shows the selection frequency of these wavenumbers. Table 3.11 shows the bands that were selected 100% of times in the PLSDA model. The 781 cm^{-1} band, attributed to the ring structure of pyrimidine bases, is identified as important for delineating non-cancer from metastatic AC tissue. Other bands recognised as important for characterising include: 1003 (phenylalanine), 1152 (protein, carotenoid), 1288, 1304, 1320 (collagen), 1338 (amide III), 1445 (phospholipid, protein), 1527 (carotenoid), 1551 (tryptophan), 1581 (nucleic acid), 1604, 1617 (amino acids), and 1656 cm^{-1} (amide I).

Table 3.11- Raman shifts selected 100% of times from the VIP.

Bands selected for classification 100% of the time	
Band Wavelength	**Tentative assignment**
781.2-783.3	Pyrimidine base
1003	Phe
1152	Protein/carotenoid
1288	Collagen
1294-1338	Lipid/collagen, Amide III
1429-1460	Protein, phospholipid
1536-1600	Nucleic acid, amino acid
1603	Amino acid
1613-1620	Amino acid
1654-1666	Amide I
1670-1672	Amide I
1676-1677	Amide I

The VIP plot was compared to the analysis of the PLSDA weights in order to identify the biochemicals that were identified as important and which pathology they are characteristic of. It was found that the peaks more indicative of metastatic AC LN measurements include 1296 cm^{-1}, assigned to lipid/collagen, 1437 and 1456 cm^{-1} assigned to phospholipid/protein. Whilst those more indicative of non-cancer LN tissue are peaks at 781 cm^{-1}, assigned as pyrimidine, 1152 cm^{-1},assigned as protein/carotenoid, 1294 cm^{-1},assigned as cytosine, 1559 cm^{-1}, assigned as tryptophan and 1576 cm^{-1} assigned as nucleic acid.

3.1.2 Raman Spectroscopy of Femoral Lymph Nodes

This study also measured femoral lymph nodes to assess the use of Raman spectroscopy for differentiation between non-cancer samples and those with metastatic squamous cell carcinoma from cervical cancers. Femoral lymph nodes are found in the groin and often referred to as inguinal nodes.

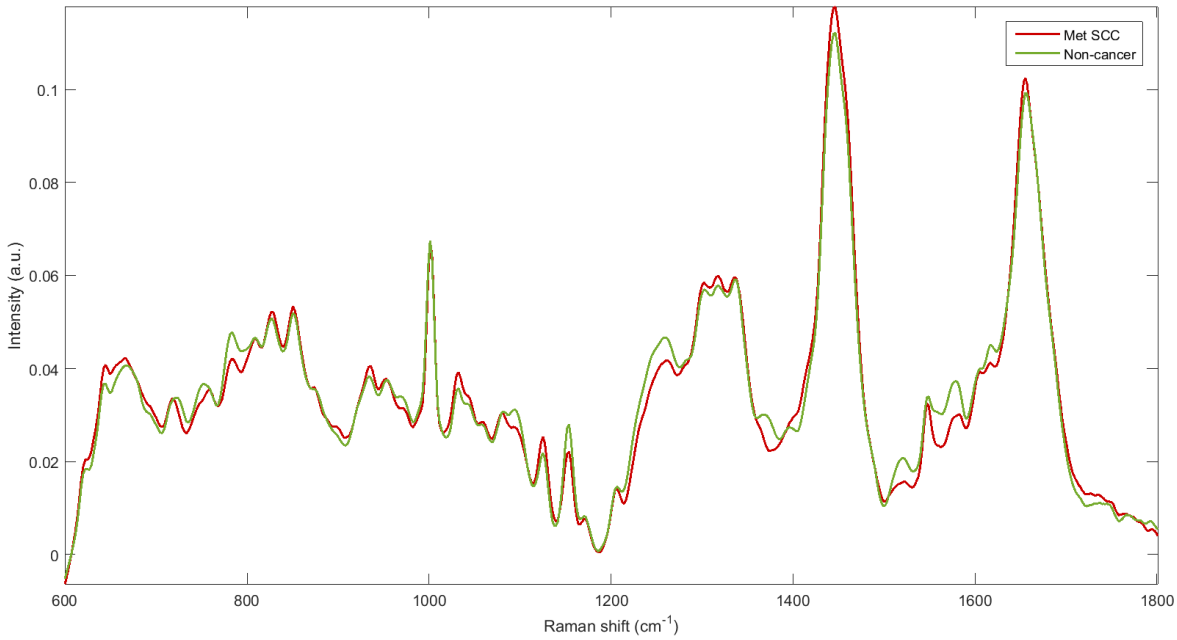

Figure 3.39- Mean spectra of metastatic SCC and non-cancer femoral lymph node measurements from 120 and 160 acquisitions respectively.

Figure 3.39 shows the resulting mean spectra from measurements of the metastatic and non-cancer femoral lymph nodes. Similarly to the oesophageal LN data, differences are mainly present as a result of intensity variation as oppose to changes in peak position. Visual differences appear most noticeably at 784 cm^{-1}, assigned as pyrimidine, 1261 cm^{-1}, assigned as amide III, 1365 cm^{-1}, assigned as tryptophan, 1527 cm^{-1}, assigned as carotenoid and 1581 and 1617 cm^{-1}, assigned as amino acid, where they appear less intense in the metastatic mean spectrum than the non-cancer.

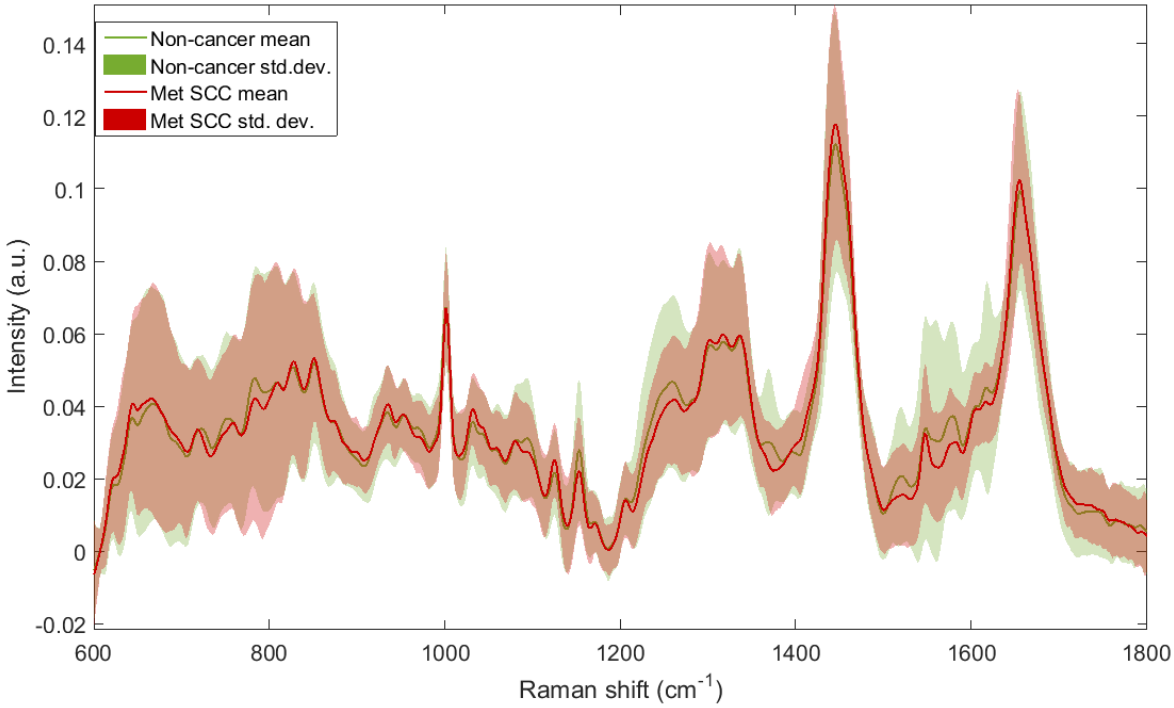

Figure 3.40- Standard deviation plotted around the mean spectra of metastatic SCC and non-cancer femoral lymph node measurements from 120 and 160 acquisitions respectively.

To evaluate the variation of the dataset the standard deviation was plotted around the mean, displayed in Figure 3.40. It can be seen that variation is higher in both the groups, mainly in the DNA regions of the spectra between 600–900 cm^{-1} and the amide III peak and lipid region of 1200–1400 cm^{-1}. The level of variation seems to be similar across both the metastatic and non-cancer LN measurements except at the amide III peak (1261 cm^{-1}) and the nucleic and amino acid peaks around 1500-1620 cm^{-1}. This could be due to a higher number of samples, and hence measurements, in the non-cancer LN group leading to a higher chance of variation within the dataset.

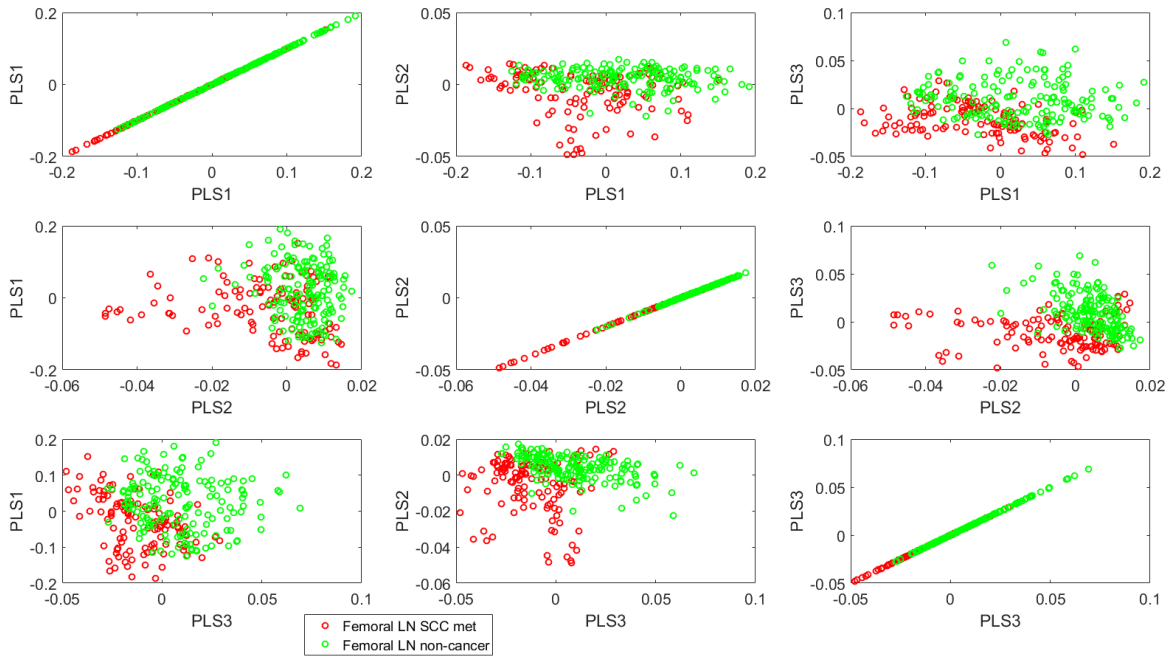

Figure 3.41- PLS scores plot of the metastatic SCC and non-cancer femoral lymph node measurements.

The subtlety in differences between the metastatic and non-cancer measurements requires multivariate analysis to identify significant variations between the two groups and cross-validation to evaluate the ability for classification. PLSDA was performed and cross-validated using leave-one-sample-out. Through a majority wins approach, a sensitivity and specificity of 83% and 75% respectively for the identification of metastatic SCC in femoral lymph nodes was calculated. The confusion matrix for the validation is seen in Table 3.12.

Table 3.12- Confusion matrix for the classification of metastatic SCC in femoral lymph nodes.

		True group	
		Metastatic SCC	**Non-cancer**
Predicted group	**Metastatic SCC**	5	2
	Non-cancer	1	6
Total LN samples		6	8
Sensitivity		0.8333	0.7500
Specificity		0.7500	0.8333

The scores plot displayed in Figure 3.41 shows the clustering of the two datasets. The scores for PLS components 1, 2 and 3 are generally more negative in the metastatic group and more positive in the non-cancer group. The PLSDA weights for the first three components are presented in Figure 3.42 and Table 3.13 lists the main negative and positive peaks with their tentatively assigned biochemicals.

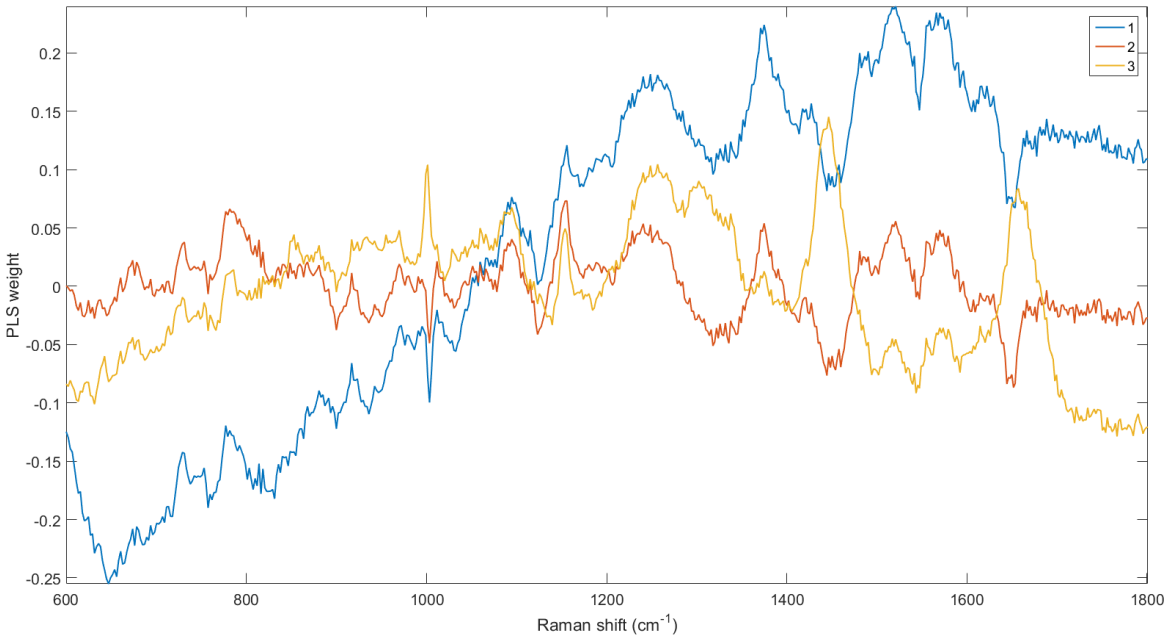

Figure 3.42- PLS weight spectra of the first three PLS component.

The peaks in the negative direction are more representative of the scores that are also negative, and thus metastatic LN measurements. Those that are in the positive direction are more characteristic of scores also positive, and therefore non-cancer LN measurements for the first three components. Negative peaks include 899.7 cm^{-1}, assigned as adenine, 935.9 cm^{-1}, assigned as protein, 1003 and 1032 cm^{-1}, assigned as phenylalanine, 1123, 1445 and 1078 cm^{-1}, assigned as lipid peaks, 1318 cm^{-1}, assigned as collagen and 1654 cm^{-1}, assigned as amide I. Positive peaks, thus those that are more indicative of non-cancer measurements, include 728, 781, 1094 and 1480 cm^{-1}, assigned as DNA, 753, 853, 1375 and 1559 cm^{-1}, assigned as amino acid, 1521 cm^{-1}, assigned as carotenoid and 1570 cm^{-1}, assigned as nucleic acid. Component three slightly conflicts with the other two as the phenylalanine peak is positive, but with a small

shift in wavenumber at 1001 cm^{-1} as opposed to 1003 cm^{-1}. This might suggest that the phenylalanine is slightly altered, perhaps by the microenvironment, in metastatic tissue causing a shift in the wavelength of the scattered light. Furthermore, a peak assigned as collagen/lipid at 1302 cm^{-1} is present in the positive direction. However, in the other two components the lipid and collagen peaks appear to be present more often in the negative direction.

Table 3.13- Main peaks and their biochemical assignments from the PLS component weights.

Component 1				**Component 2**				**Component 3**			
-ve band		**+ve band**		**-ve band**		**+ve band**		**-ve band**		**+ve band**	
900	Adenine	729	DNA	900	Adenine	674	Cystine	900	Adenine	642	Tyr
936	Protein	753	Tryp	936	Protein	731	DNA			674	Cystine
1003	Phe	777	Pyrimidine	951	Hydroxyapatite, cholesterol	781	Pyrimidine			786	Pyrimidine
1032	Phe	1094	DNA	1003	Phe	1094	DNA			853	Tyr
1123	Lipid, protein	1156	Protein, carotenoid	1032	Phe	1156	Protein, carotenoid			1001	Phe
1445	Phospholipid, protein	1257	Carb, amide III	1078	Lipid, protein	1249	Carb, amide III			1094	DNA
1654	Amide I	1375	Tryp	1123	Lipid, protein	1365	Tryp			1154	Protein, carotenoid
		1490	DNA	1206	Protein	1481	DNA			1257	Carb, amide III
		1521	Carotenoid	1318	Collagen	1521	Carotenoid			1302	Collagen, lipid
		1559	Tryp	1445	Phospholipid, protein	1570	Nucleic acid			1375	Tryp
		1617	Amino acid	1652	Amide I					1446	Phospholipid, protein
										1517	Carotenoid
										1574	Nucleic acid
										1657	Amide I

Identification of the important peaks for discrimination by VIP, suggests that the wavenumbers where amino acids, nucleic acids, collagen, amide I and

III, lipid and carotenoid peaks are useful for delineation of metastatic SCC from non-cancer LN measurements. The VIP bar plot is shown in Figure 3.43 and the wavenumbers with 100% selection are listen in Table 3.14 with the tentative assignments for the peaks present within the bands. Through comparison of the VIP with the PLSDA weights it can be seen that metastatic SCC LN can be identified from non-cancer LN measurements from a higher intensity of peaks assigned as lipid and collagen. This also indicates a higher content of these biochemicals within the tissue. Moreover, metastatic LN also appears to have less intense peaks assigned as DNA, nucleic acids, amino acid and amide III than non-cancer LN measurements.

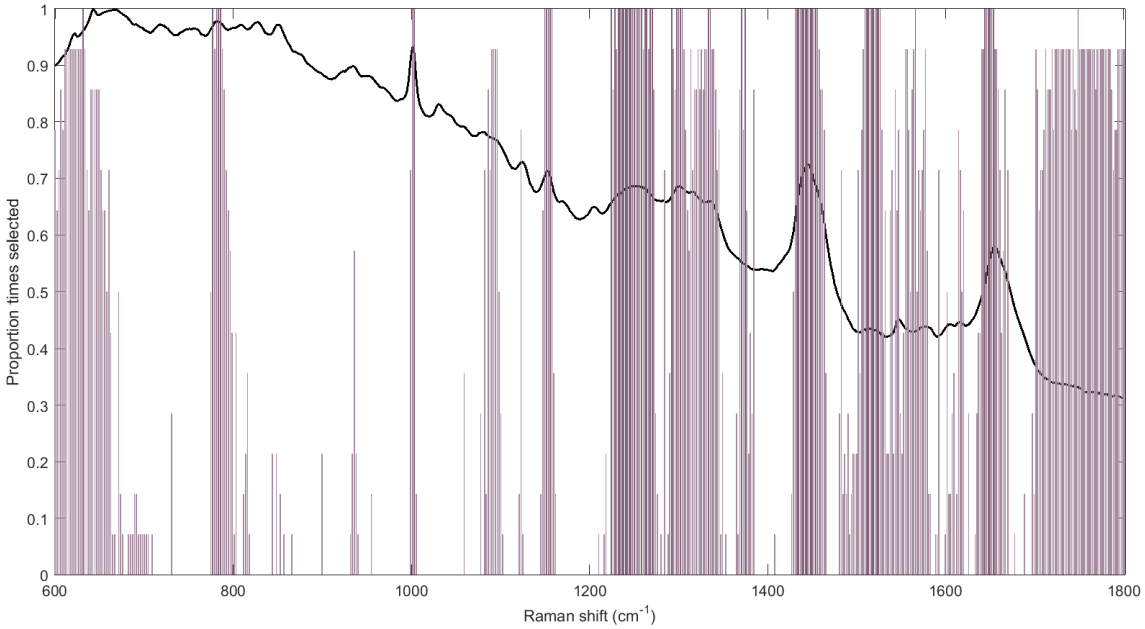

Figure 3.43- VIP bar plot overlaid onto a mean spectrum of the dataset to show the proportion of times specific Raman bands were included in classification.

Table 3.14- Table of the Raman shifts that were selected 100% of the time for classification with their biochemical assignments.

Bands selected for classification 100% of the time	
Band Wavelength	**Tentative assignment**
631.1	Amino acid
776.9-781.3	Pyrimidine base
1001-1003	Phe
1150-1158	Carotenoid, protein
1224	Protein, amide III, carb
1228	Protein, amide III, carb
1233-1258	Protein, amide III, carb
1262-1264	Amide III, carb
1268-1270	Amide III, carb
1292	Collagen
1298-1304	Collagen, lipid
1334-1336	Amide III
1371	Tryp
1375	Tryp
1431-1456	Protein, phospholipid
1509-1527	Carotenoid
1644-1654	Amide I
1749	Lipid

3.1.3 Raman Spectroscopy of Head and Neck Lymph Nodes

The averaged spectra for measurements from metastatic squamous cell carcinoma (SCC) and non-cancer lymph nodes of the head and neck are displayed in Figure 3.44. Similarly to the oesophageal lymph node data the main differences between the spectra are peak intensities as oppose to position. Visual differences include less intense peaks assigned as DNA and nucleic acid and more intense peaks assigned as lipid, amide I and amide III in the metastatic data.

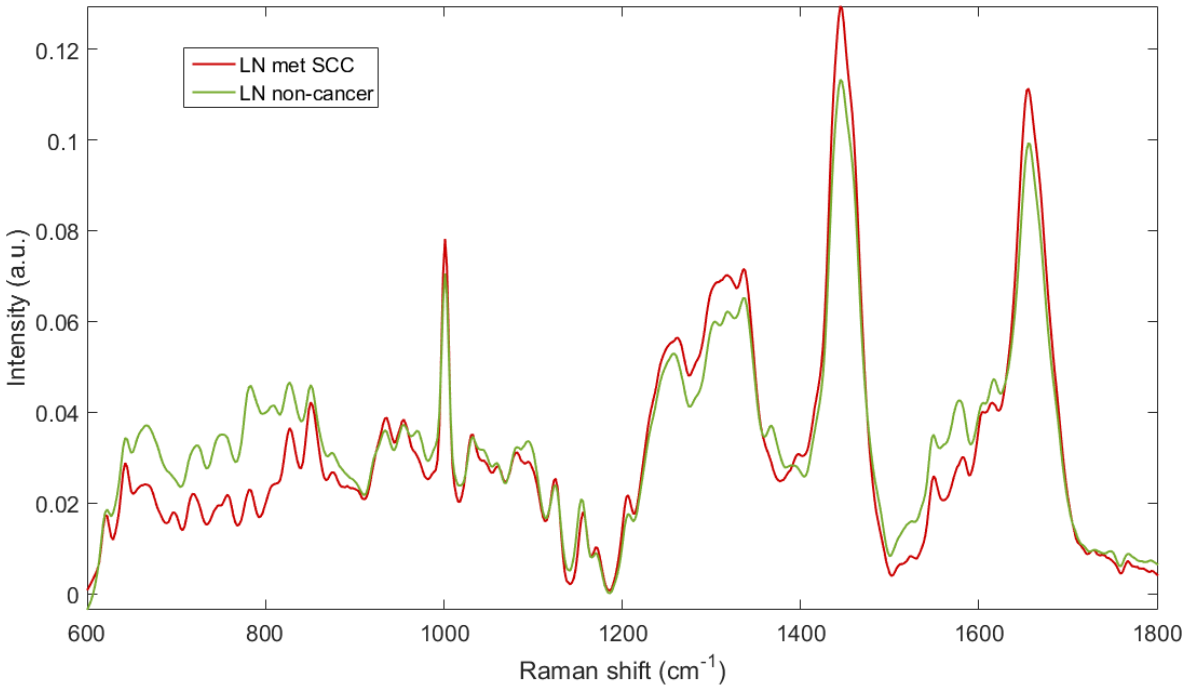

Figure 3.44- Mean spectra of metastatic SCC and non-cancer data from head and neck lymph nodes averaged across 100 and 260 measurements respectively.

The standard deviation was plotted around the averaged spectra and shown in Figure 3.45 to show the variation of the dataset. The non-cancer LN data appear to exhibit much more variability than the metastatic SCC data, unlike the oesophageal dataset. However, in this dataset there are more non-cancer samples included than metastatic samples which may lead to this increase in deviation. The peaks characteristic of DNA and amino acids are still typically

lower and those characteristic of lipids higher in the metastatic measurements even in the standard deviation.

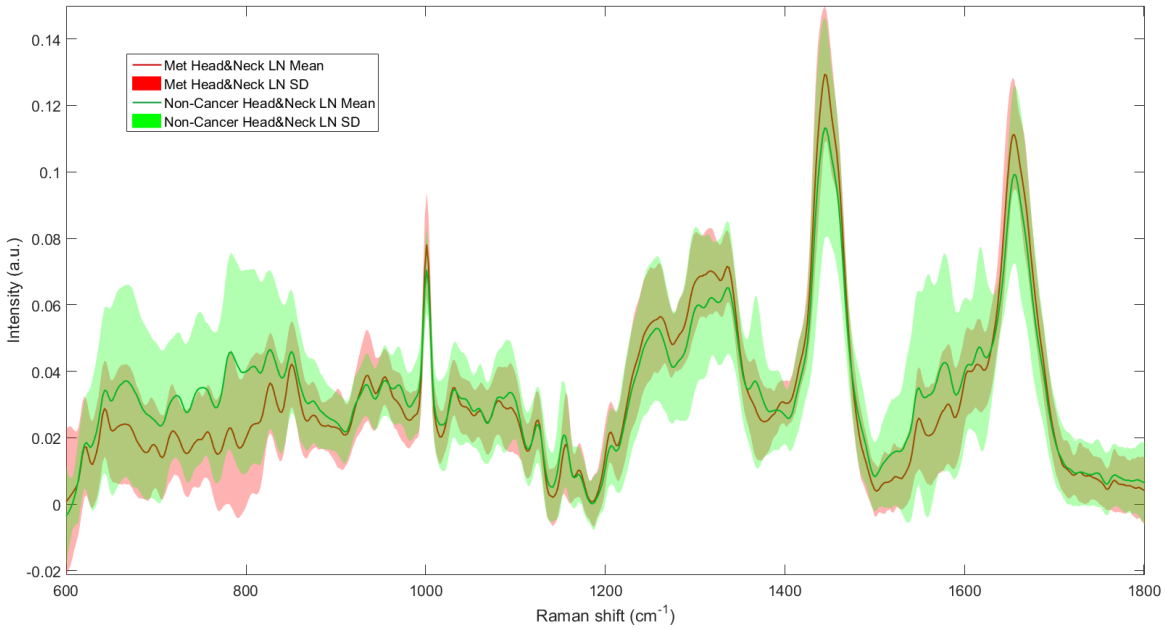

Figure 3.45- Standard deviation plotted around the mean spectra of SCC metastatic and non-cancer data from head and neck lymph nodes averaged across 100 and 260 measurements respectively.

Multi-variate analysis was carried out to assess if the two groups could be distinguished from one another analytically and also to reveal the main wavenumber regions in the spectra for this discrimination. PLSDA, a supervised technique was employed and a leave-one-sample-out cross validation used to quantify the classification capabilities in the form of sensitivities and specificities using a majority wins approach.

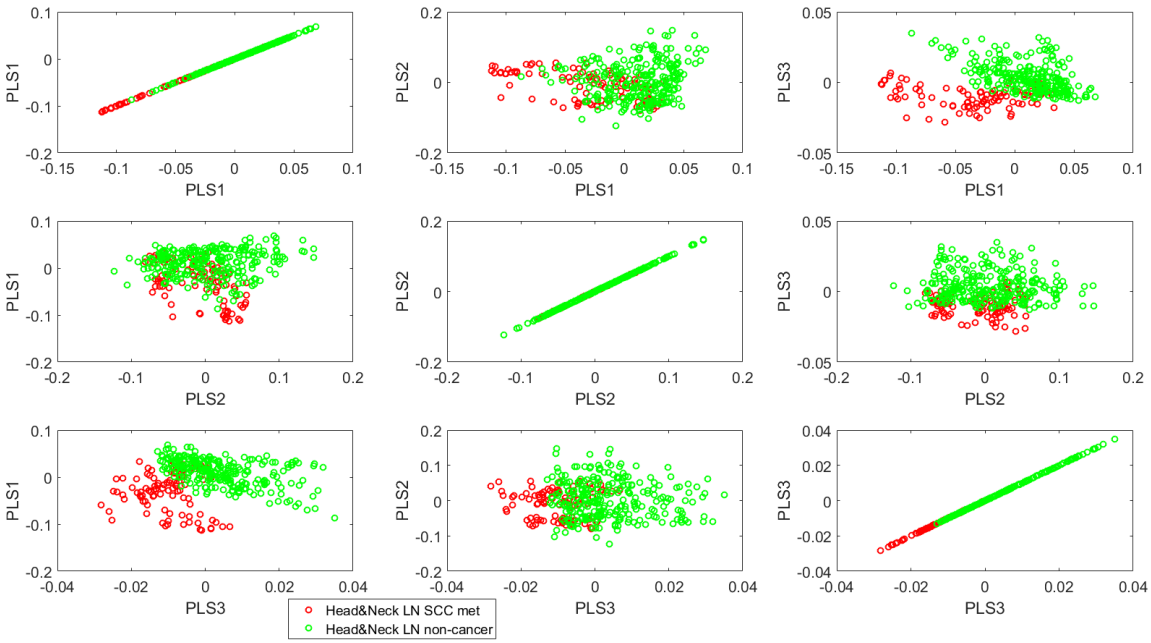

Figure 3.46- PLSDA scores plot showing clustering of the metastatic and non-cancer measurements from head and neck lymph nodes.

The PLSDA scores plots shown in Figure 3.46 displays the dataset scores depending on the first three PLS components. The scores plot enables visualisation of the dataset clustering. It is seen that the metastatic and non-cancer lymph node measurements typically cluster with others from the same group with a slight amount of overlap. PLS component 1 has mainly positive scores pertaining to the non-cancer LN group and negative belonging to the metastatic SCC LN group. In PLS component 2 it is more difficult to discern between the groups, as the non-cancer scores are spread widely across both the negative and positive directions on the axis. However, metastatic scores are more frequently negative than positive. PLS component 3 is seen to separate primarily negative metastatic scores from the more positive non-cancer LN scores.

Table 3.15- Confusion matrix for the classification of Metastatic SCC in head and neck lymph nodes.

		Actual group	
		Metastatic SCC	**Reactive**
Predicted group	**Metastatic SCC**	4	1
	Reactive	1	12
Total LN samples		5	13
Sensitivity		0.8000	0.9231
Specificity		0.9231	0.8000

A leave-one-sample-out cross-validation resulted in a sensitivity of 80% and specificity of 92% for the identification of metastatic SCC LN from non-cancer LN tissue, when employing a majority wins approach (Table 3.15).

The PLSDA weights from the first three components, displayed in Figure 3.47, show the peaks that contribute to the component, and the direction in which they are present indicate whether they are more indicative of the negative or positive scores. The main peaks from the weights spectra of the first three PLSDA components are listed in Table 3.16 and show the tentative biochemical assignments.

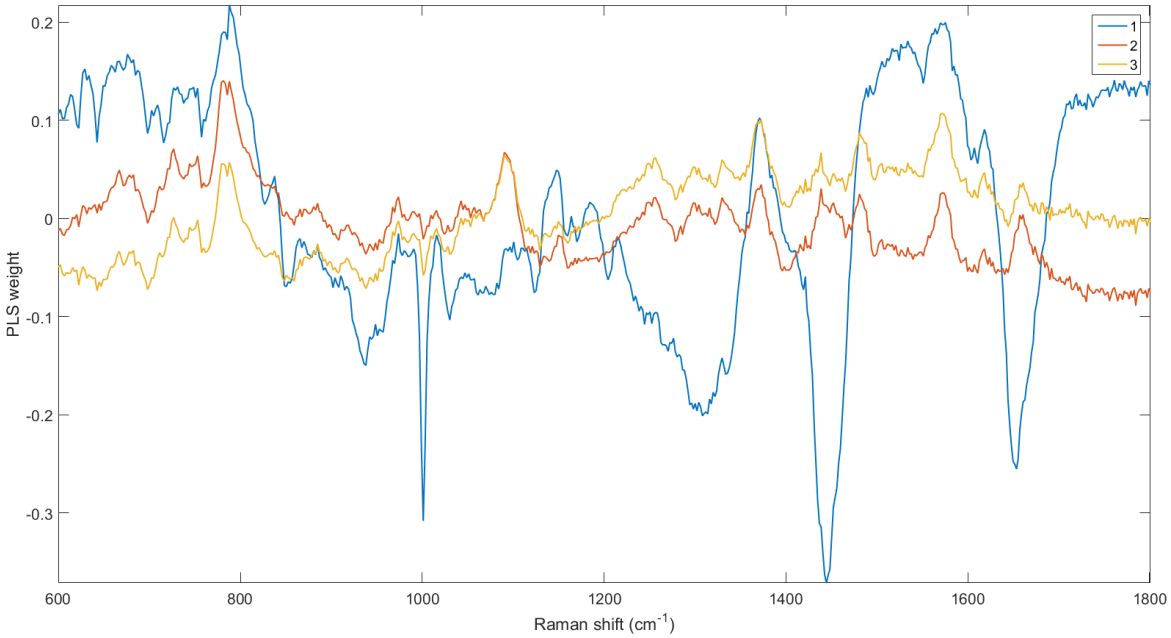

Figure 3.47- Weight spectra of the top three components in PLSDA.

Table 3.16- Main peaks from PLS weights with tentative biochemical assignments.

Component 1				Component 2				Component 3			
-ve band		+ve band		-ve band		+ve band		-ve band		+ve band	
848.4	Tyr	626.7	Phe	697.9	Cystine	675.8	Amino acid	697.9	Cystine	675.8	Amino acid
938	Protein	666.9	Amino acid	767	Amino acid	781.3	Pyrimidine	767	Amino acid	781.3	Pyrimidine
954.9	Hydroxyapatite, cholesterol	709	DNA	938	Protein	1090	DNA	938	Protein	1090	DNA
1001	Phe	726.6	DNA	1001	Phe	1257	Carbohydrate, amide III	1001	Phe	1257	Carbohydrate, amide III
1030	Phe	748.5	DNA	1278	Collagen	1300	Lipid/collagen	1278	Collagen	1300	Lipid/collagen
1061	Lipid	781.3	Pyrimidine			1330	Amide III			1330	Amide III
1080	Protein, lipid	787.8	Pyrimidine			1373	Tryp			1373	Tryp
1123	Lipid, protein	1148	Protein, carotenoid			1439	Phospholipid, protein			1439	Phospholipid, protein
1204	Protein	1373	Tryp			1481	DNA			1481	DNA
1308	Lipid, collagen	1492	DNA			1574	Nucleic acid			1574	Nucleic acid
1440	Phospholipid, protein	1517	Carotenoid			1618	Amino acid			1618	Amino acid
1654	Amide I	1534	Tryp			1661	Amide I			1661	Amide I
		1570	Nucleic acid								
		1618	Amino acid								

The negative peaks from the three components are more representative of the negative scores, and thus metastatic SCC measurements. These include peaks at 698 and 848 cm^{-1} assigned as the amino acids cysteine and tyrosine, assigned protein peaks at 938 and 1204 cm^{-1}, phenylalanine at 1001 and 1030 cm^{-1}, collagen at 1278 cm^{-1}, and peaks at 1061, 1080, 1123, 1308 and 1440 cm^{-1}, assigned as lipid. An amide I peak at 1654 cm^{-1} is also present negatively in the weights of component 1. The positive peaks, and thus more indicative of the positive scores which are primarily more characteristic of non-cancer LN measurements include DNA and nucleic acids, amide III and nucleic acids. Peaks assigned as DNA and nucleic acids comprise 709, 727, 749, 781, 1090, 1481

and 1574 cm^{-1}, amide III at 1257 cm^{-1} and amino acid at 1330 cm^{-1}. Conflicting peaks include a peak at 1300 cm^{-1} assigned as collagen/lipid which is positive in the weights for component 2 and 3 however negative in component 1 at 1308 cm^{-1}. Also an assigned amide I peak is positive in components 2 and 3 at 1661 cm^{-1}, yet negative in component 1 at 1654 cm^{-1} and a peak at 1440 cm^{-1}, assigned as phospholipid/protein, is negative in component 1 and positive in components 2 and 3 at 1439 cm^{-1}. These discrepancies between components could be highlighting a difference in the Raman shift of the collagen and amide I peaks rather than an alteration between the intensities of the peaks.

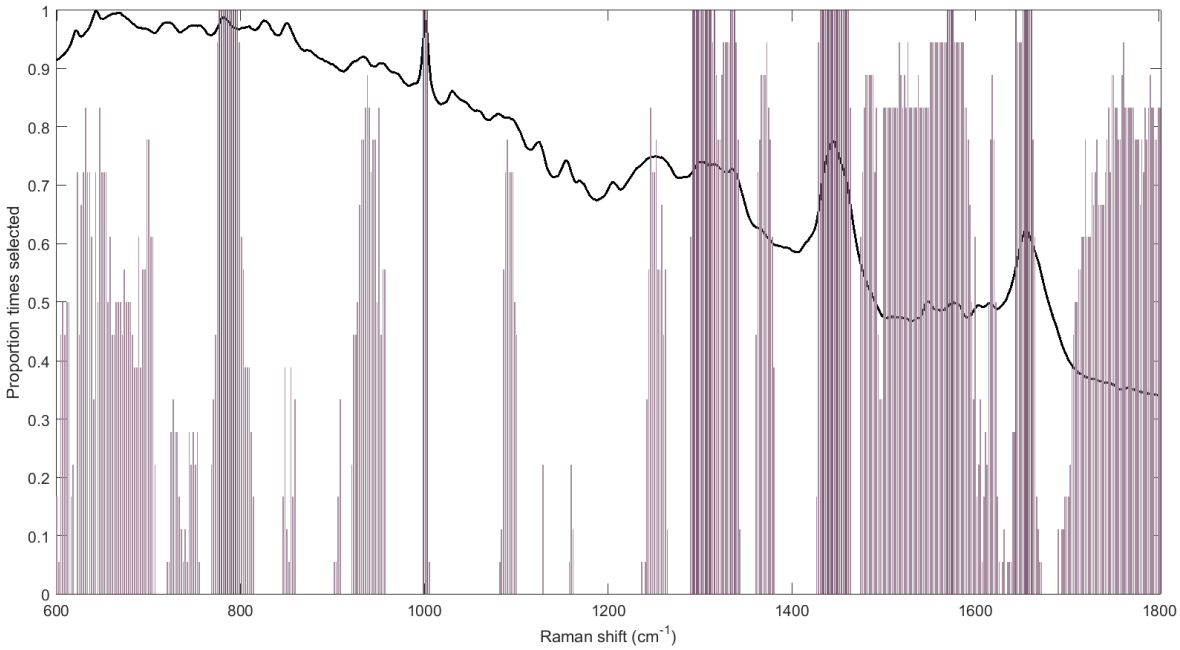

Figure 3.48- Bar chart of the VIP results plotted in front of the mean spectrum showing the wavelengths and the proportion of times they were selected for classification.

VIP was carried out to identify the important peaks for classification, shown as a bar chart in Figure 3.48. Table 3.17 displays the wavelengths that were used 100% of the time for helping discriminate between the two groups, these include peaks assigned as DNA components, phenylalanine, lipid, collagen and amide I and III. Comparison of the VIP results and the PLSDA weights enabled the recognition of peaks contributing most to variation, and which group they are indicative of. A peak at 781 cm^{-1}, assigned as pyrimidine, is supposedly more indicative of non-cancer LN measurements and was found to be important for

classification of metastatic nodes as the peak appears at a lower intensity in these samples' spectra. This is also seen for peaks at 1300, 1439, 1574 and 1661 cm^{-1}, assigned as lipid/collagen, amide III, phospholipid/protein, nucleic acids and amide I respectively. These are all present at lower intensities in the metastatic SCC measurements compared to the non-cancer LN measurements. Peaks identified as important by VIP and are supposedly more characteristic of metastatic SCC include 1001, 1278, 1308, 1440 and 1654 cm^{-1} assigned as phenylalanine, collagen/lipid, collagen, phospholipid/protein and amide I respectively. Again, the phospholipid/protein and amide I peak could represent a change in Raman shift rather than intensity.

Table 3.17- Table of Raman bands selected 100% for classification identified through VIP along with their biochemical assignments.

Bands selected for classification 100% of the time	
Band Wavelength	**Tentative assignment**
776.9-796.5	Pyrimidine base
999-1003	Phenylalanine
1292-1312	Lipid/collagen, Amide III
1316	Collagen
1334-1338	Amide III
1431-1462	Protein, phospholipid
1570-1575	Nucleic acid
1644	Amide I
1652-1661	Amide I

Point measurements from all the three LN regions were acquired, after the oesophageal lymph node samples were found to be largely affected by carbon particles, to keep continuity and to enable the comparison of the measurements. However, maps were also acquired from the head and neck nodes in order to evaluate if there was a loss in performance and thus classification accuracy when only analysing point measurements. The results from the maps were analysed in the same way as the point measurements, so that a comparison could be made. Figure 3.49 shows the non-baseline corrected averaged spectra of the metastatic SCC and non-cancer LN Raman maps. Visually the spectra are alike to that of the point data in Figure 3.44 and the differences between the two LN groups are similar. Such as, in the mean metastatic SCC spectrum, less intense peaks in the DNA region of 600-900 cm^{-1}, in the nucleic acid and carotenoid regions between 1500-1620 cm^{-1}, and more intense phenylalanine, amide, collagen and lipid peaks between 1000-1440 cm^{-1} and 1654 cm^{-1}.

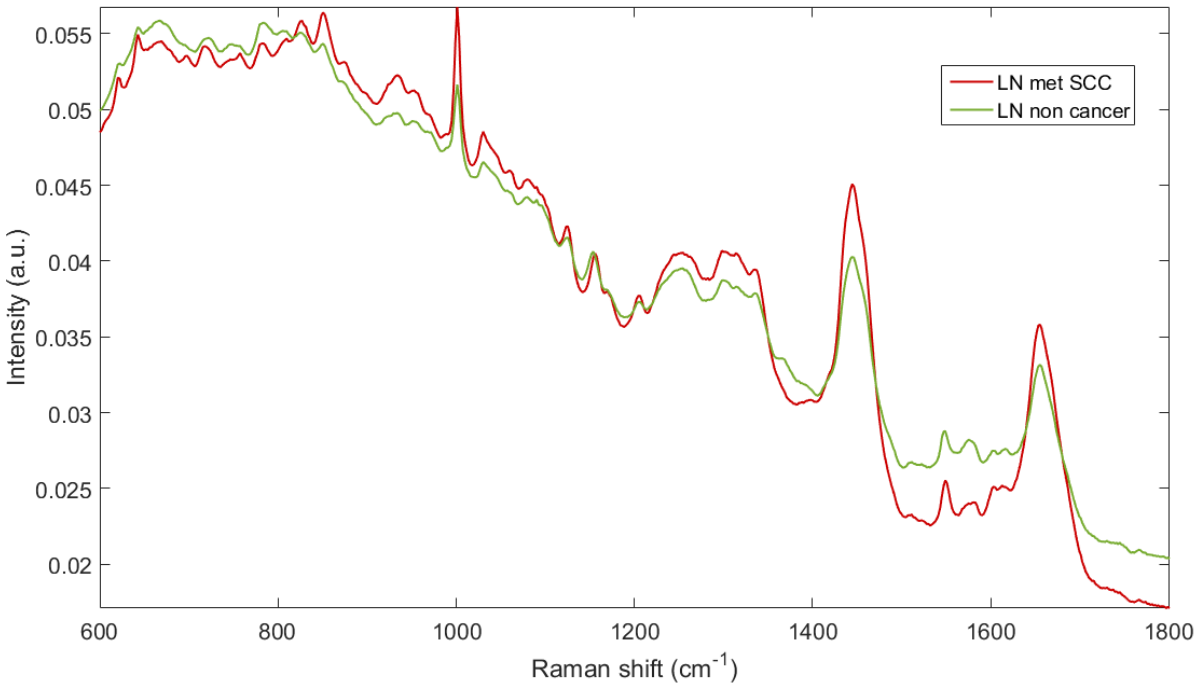

Figure 3.49- Mean spectra for metastatic SCC and non-cancer head and neck lymph node measurements from Raman maps resulting in 44802 and 131738 spectral measurements.

PLSDA was carried out on the data for comparison with the point data. The scores plot is shown in Figure 3.50 and resembles that of the point measurements in Figure 3.46 where the scores for the non-cancer

measurements are more frequently positive whilst that of the metastatic SCC are more often negative in the first three components.

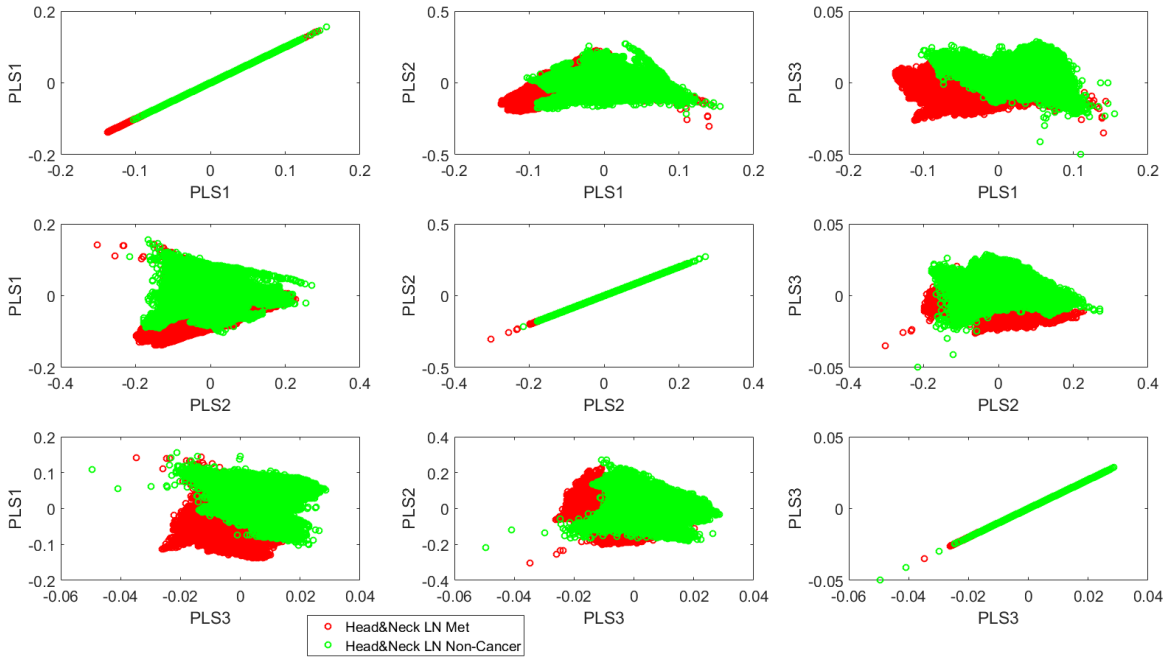

Figure 3.50- PLSDA scores plots for metastatic SCC and non-cancer head and neck LN measurements for the first three components.

Cross validation was also performed and produced a classification sensitivity of 80% and a specificity of 92% for the identification of metastatic cancers using a majority wins approach for spectra per sample, the same results calculated in the point spectra dataset. The confusion matrix for this analysis is shown in Table 3.18.

Table 3.18- Confusion matrix for classification of metastatic SCC in head and neck LN.

		True group	
		Metastatic SCC	**Non-cancer**
Predicted group	**Metastatic SCC**	4	1
	Non-cancer	1	12
Total LN samples		5	13
Sensitivity		0.8000	0.9231
Specificity		0.9231	0.8000

The spectra for the weights of the first three components are displayed in Figure 3.51 and are also similar to that of the point spectra data set in Figure 3.47. The main peaks listed in Table 3.19 also align with those of the point spectra dataset in Table 3.16, such as positive assigned DNA and nucleic acid peaks at 727, 748, 779, 792, 1090, 1485 and 1575 cm^{-1}, amino acids at 667, 675, 707, 1188, 1617 cm^{-1} and amide III at 1260 and 1334 cm^{-1}. Negative peaks assigned as collagen are present at 876 and 1314 cm^{-1}, lipid at 1059, 1082 and 1445 cm^{-1}, and phenylalanine at 1001 cm^{-1}. Similarly, the phospholipid/protein peak is present in the positive weights but at a lower Raman shift of 1439 cm^{-1} and collagen at 1300 cm^{-1} but again at a lower Raman shift.

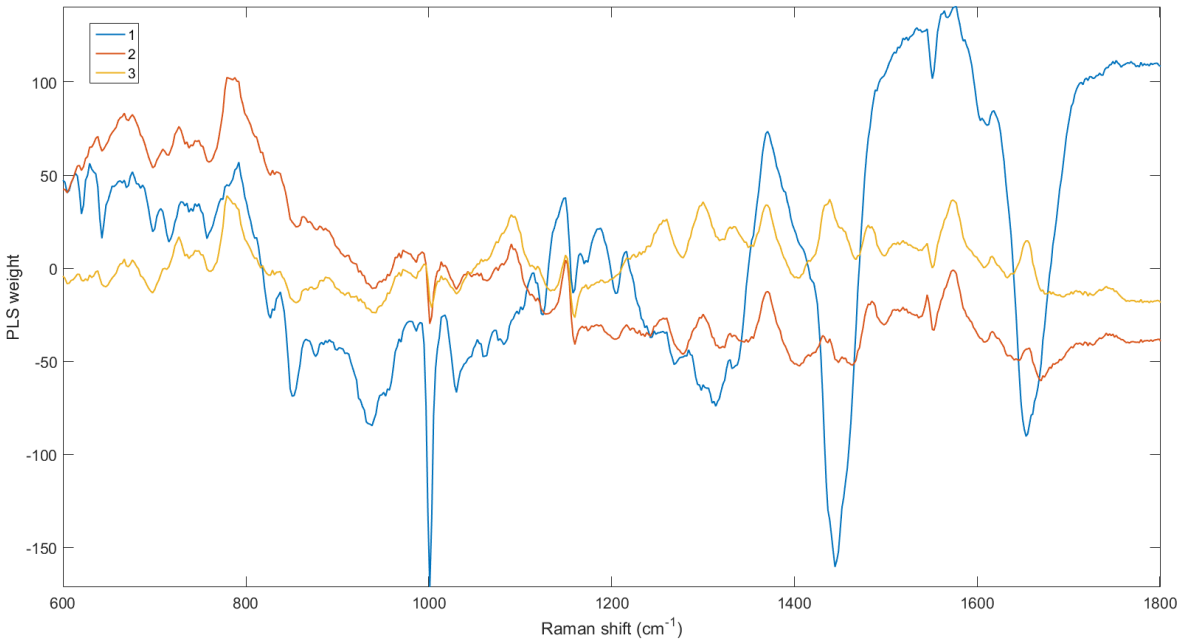

Figure 3.51- PLSDA weight spectra of the first three components.

Table 3.19- Table of the main peaks from the PLSDA weights with biochemical assignments.

Component 1				**Component 2**				**Component 3**			
-ve band		**+ve band**		**-ve band**		**+ve band**		**-ve band**		**+ve band**	
827	Nucleic acids	667	Amino acid	857	Tyr	667	Amino acid	855	Tyr	667	Amino acid
851	Tyr	676	Amino acid	940	Protein	727	DNA	942	Protein	727	DNA
876	Collagen	707	Amino acid	1001	Phe	779	Pyrimidine	1001	Phe	779	Pyrimidine
938	Protein	726.6	DNA			1090	DNA			1090	DNA
1001	Phe	748.5	DNA			1150	Protein/ carotenoid			1150	Protein/ carotenoid
1030	Phe	792.2	Pyrimidine			1261	Carb/ amide III			1261	Carb/ amide III
1059	Lipid	1150	Protein, carotenoid			1300	Lipid, collagen			1300	Lipid, collagen
1082	Lipid, protein	1188	Amino acid			1330	Amide III			1334	Amide III
1125	Lipid, protein	1216	Protein			1369	Tryp			1369	Tryp
1204	Protein	1371	Tryp			1431	Protein, phospholipid			1439	Protein, phospholipid
1314	Collagen					1485	DNA			1481	DNA
1445	Phospholipid, protein					1545	Tryp			1575	Nucleic acid
1654	Amide I					1575	Nucleic acid			1617	Amino acid
						1617	Amino acid			1655	Amide I
						1655	Amide I				

VIP was again performed to compare the results to the point spectra dataset. The plot displayed in Figure 3.52 shows the regions of the spectra that were used for classification and the proportion of times. Those used 100% of times for classification are listed with their biochemical assignments in Table 3.20. Although certain spectral bands were found to be important for both the mapped dataset and the point spectra dataset, differences included no collagen bands and the inclusion of the lipid peak area at 1745 cm^{-1} in the mapped dataset.

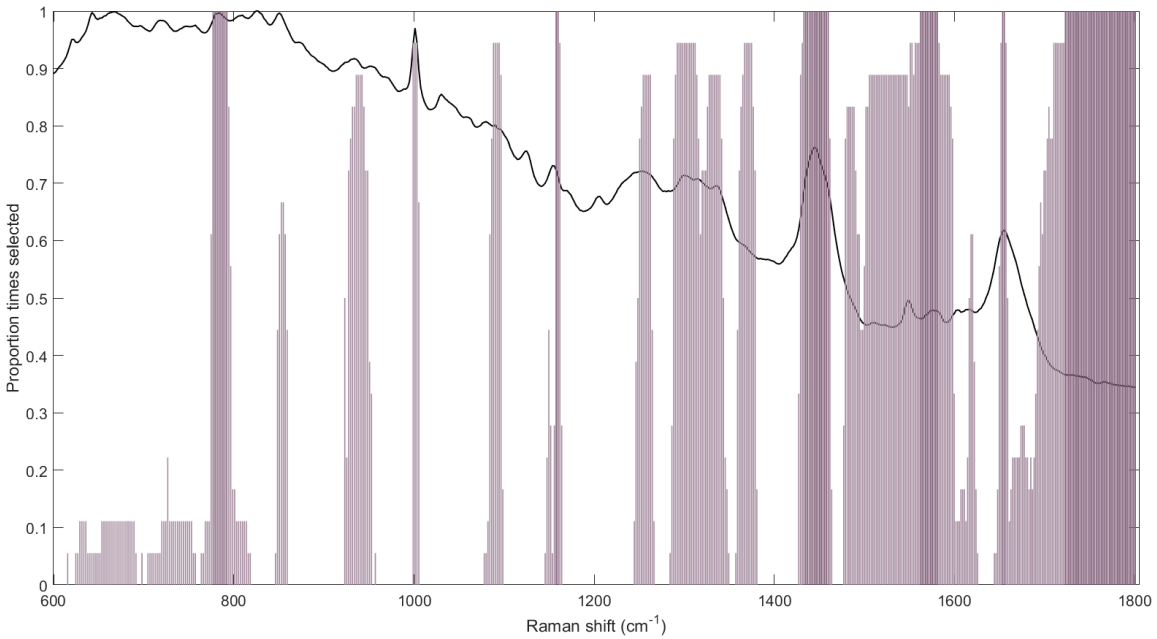

Figure 3.52- VIP plot to show the proportion of times bands were selected for classification.

Table 3.20- Table of bands selected 100% of times for classification identified through VIP along with their biochemical assignments.

Bands selected for classification 100% of the time	
Band Wavelength	**Tentative assignment**
777-792	Pyrimidine base
1158-1160	Carotenoid, protein
1433-1460	Protein, phospholipid
1562-1581	Nucleic acid
1654-1655	Amide I
1723-1801	Lipid

After the adjacent H&E sections were analysed the pathologist identified regions that were of interest. These regions were then spectrally mapped in order to create a spectral image representing biochemical content. Figure 3.53 shows the white light image of the H&E stained sample and the k-means clustered image of the adjacent mapped section. The images are labelled to show the areas of keratinocytes, SCC cells and non-cancer lymphocyte cells. Visually the areas look distinct and therefore Raman was used to assess if they could be spectrally delineated based on the biochemicals in the tissue that contribute to the scattering detected. K-means clustering was applied to create a pseudo-colour image of the map, categorised based on similar spectral components. From

visual evaluation it is clear that the regions are spectrally distinct. The keratinocytes are shown as dark blue, the SCC as cyan and the lymphocytes as red. The image shown is a 6 cluster image however, and it can be seen that there is a yellow border between the keratinocytes and the SCC and also a lighter blue and dark red within the keratinocyte region. The lymphocyte cells and the SCC cells appear to have a clear separation between the groups whereas the border between the SCC cells and the keratinocytes is less clear. Possibly as a result of diffusion of the cells between the different areas.

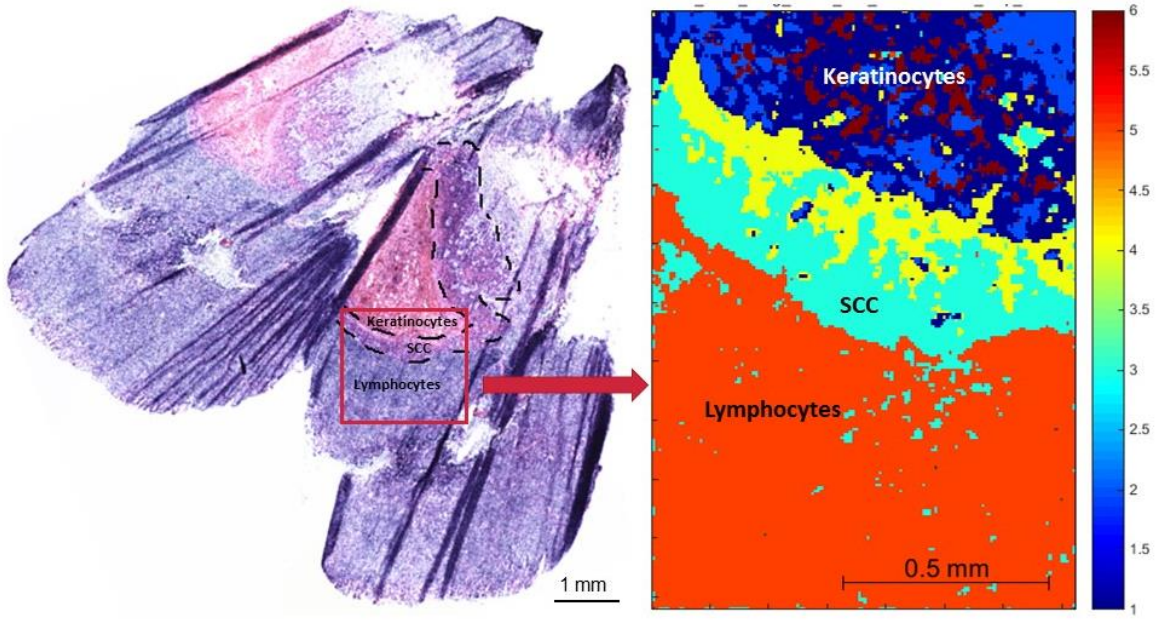

Figure 3.53- K-mean cluster image from a Raman map (right) from the area highlighted in the white light image (left).

The centroid spectra (Figure 3.54) exhibit visual differences between the measurements. Particularly in the non-cancer lymphocytes spectra where there appears to be a stronger amide III (1261 cm^{-1}) and DNA and nucleic acids (723, 784, 827, 1095 and 1581 cm^{-1}) peaks and weaker amide I (1747 cm^{-1}) and collagen (1320 cm^{-1}) assigned bands than in the SCC and keratinocyte spectra. Another variation between the spectra is the lack of a 784 cm^{-1} peak, assigned as pyrimidine base, in the keratinocyte spectrum.

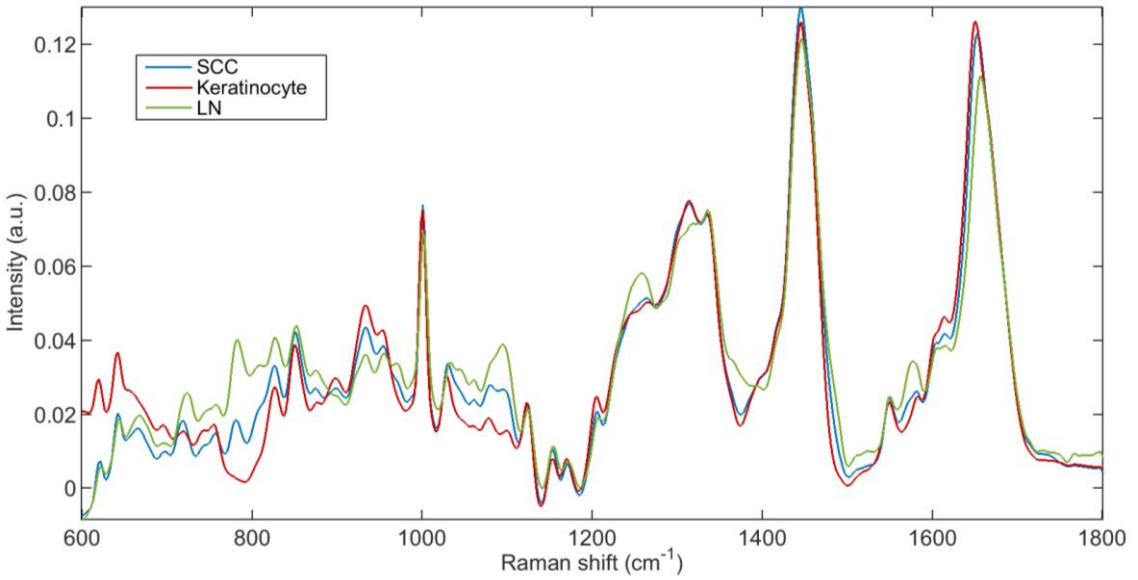

Figure 3.54- Centroid spectra from the Raman map of the SCC, keratinocyte and non-cancer lymphocyte lymph node regions.

FTIR was carried out on the same area and section of head and neck lymph node samples, to compare the spectroscopic techniques. The averaged FTIR spectra of the data taken from maps are displayed in Figure 3.55. The noise is much less in FTIR, but there are not as many spectral peaks. Visually there is a slight intensity decrease in signal around the nucleic acid assigned peaks at 1086 and 1236 cm^{-1} and around the lipid assigned peak at 1741 cm^{-1}, and a slight increase at the amide II assigned peak[230] at 1547 cm^{-1} in the metastatic averaged measurement compared to the non-cancer. PLSDA was carried out on the dataset using the FTIR data and the scores plot (Figure 3.56) showed the grouping of the data. Cross-validation carried out resulted in a sensitivity of 60% and sensitivity of 69% for the identification of head and neck lymph nodes with metastatic SCC (Table 3.21).

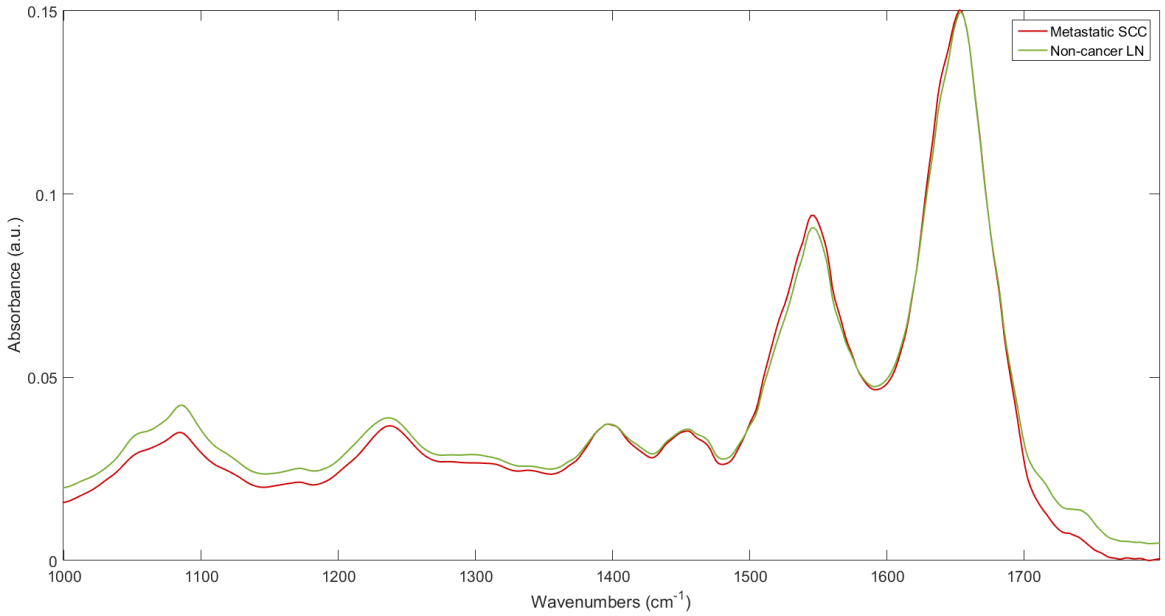

Figure 3.55- Mean FTIR spectra of metastatic SCC and non-cancer LN averaged over 3942 and 5621.

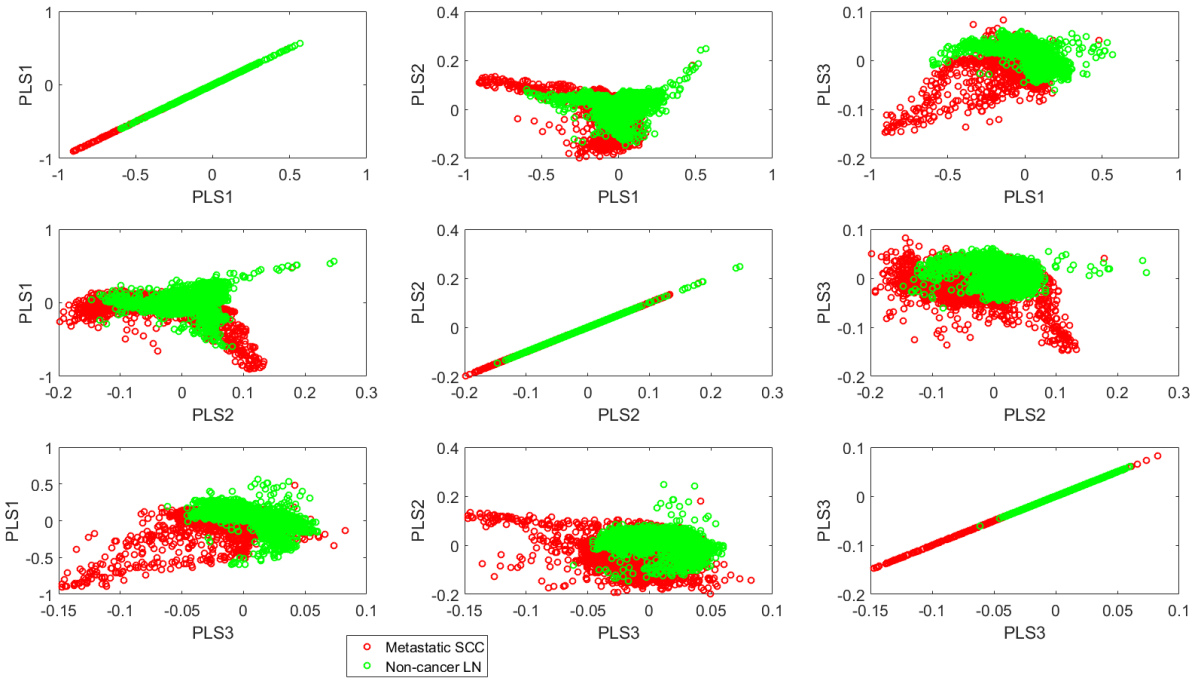

Figure 3.56- PLSDA scores plots of the top three components.

Table 3.21- Confusion matrix for the cross validation of the PLSDA model using a majority wins approach.

		True group	
		Metastatic SCC	**Non-cancer**
Predicted group	**Metastatic SCC**	3	4
	Non-cancer	2	9
Total LN samples		5	13
Sensitivity		0.6000	0.6923
Specificity		0.6923	0.6000

3.1.4 Raman Spectroscopy of Combined Lymph Nodes

In the three lymph node regions looked at the number of metastatic samples was limited, hence the ability to distinguish metastatic from non-cancer lymph nodes of the three lymph node regions combined was investigated. The mean spectra of the metastatic and non-cancer lymph node measurements are displayed in Figure 3.57 to visually assess any obvious spectral differences.

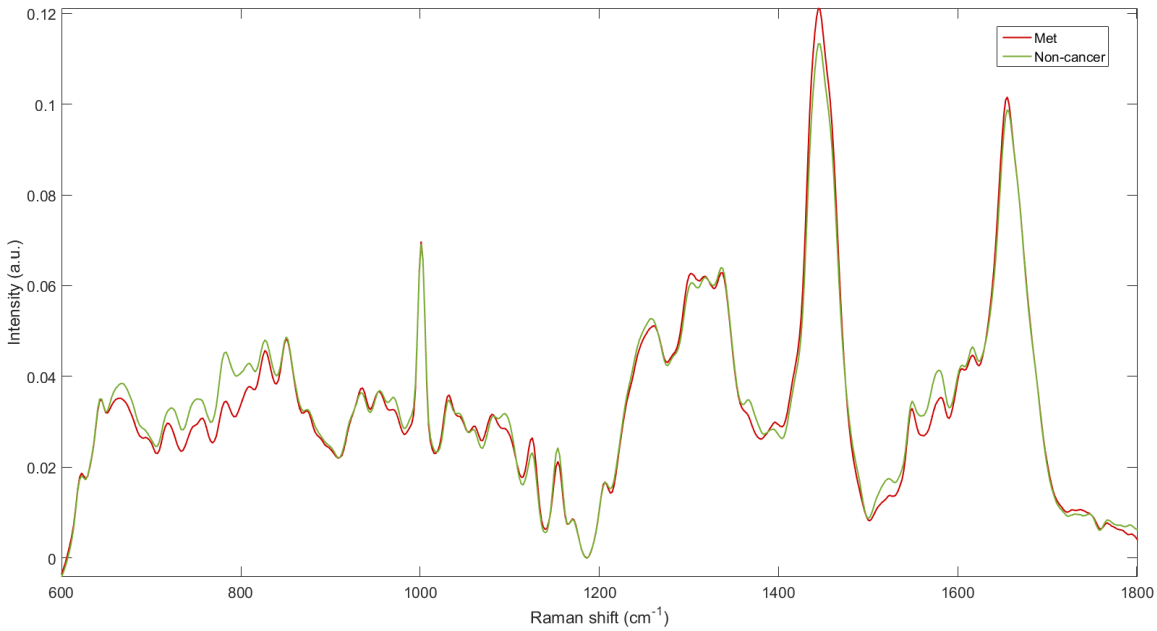

Figure 3.57- Mean spectra of metastatic (red) and non-cancer (green) lymph nodes averaged across 340 and 540 measurements respectively.

Although biochemical differences are expected between metastatic and non-cancer tissue, PCA was carried out to not only assess if separation was present based on pathology but also between data from the various lymph node regions, as the technique does not require prior information. The PCA scores plot in Figure 3.58 shows the clustering of the measurements based on the top four principal components calculated using ANOVA.

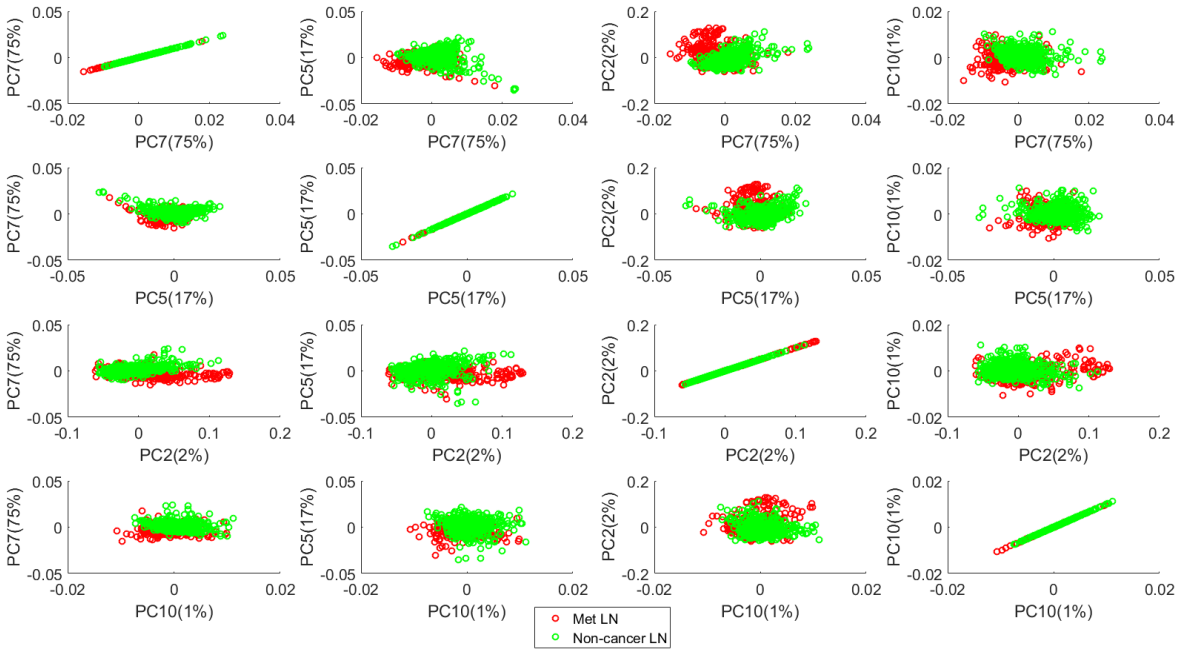

Figure 3.58-PCA plots of measurements from metastatic (red) and non-cancer (green) lymph nodes.

From ANOVA and through visually assessing the scores plot it is clear that the main component for separation of the metastatic and non-cancer LN data groups is PC7. It can be seen that in PC7 the metastatic LN data congregate mainly on the negative axis whilst the non-cancer LN data are mainly plotted on the positive. The separation for the next three components is not visually as clear. However, looking at the average and median values for the dataset, it can be seen that for PC5 the metastatic scores are generally negative whilst the non-cancer scores are positive. For PC2, the metastatic scores seem to be both positive and negative and are seen to be dispersed across the whole axis, whilst the non-cancer visually congregate closer to 0. For PC10, the metastatic are primarily negative whilst the non-cancer are slightly more positive. The loadings

of the components are displayed in Figure 3.59 and the main peaks and their assignments are listed in Table 3.22.

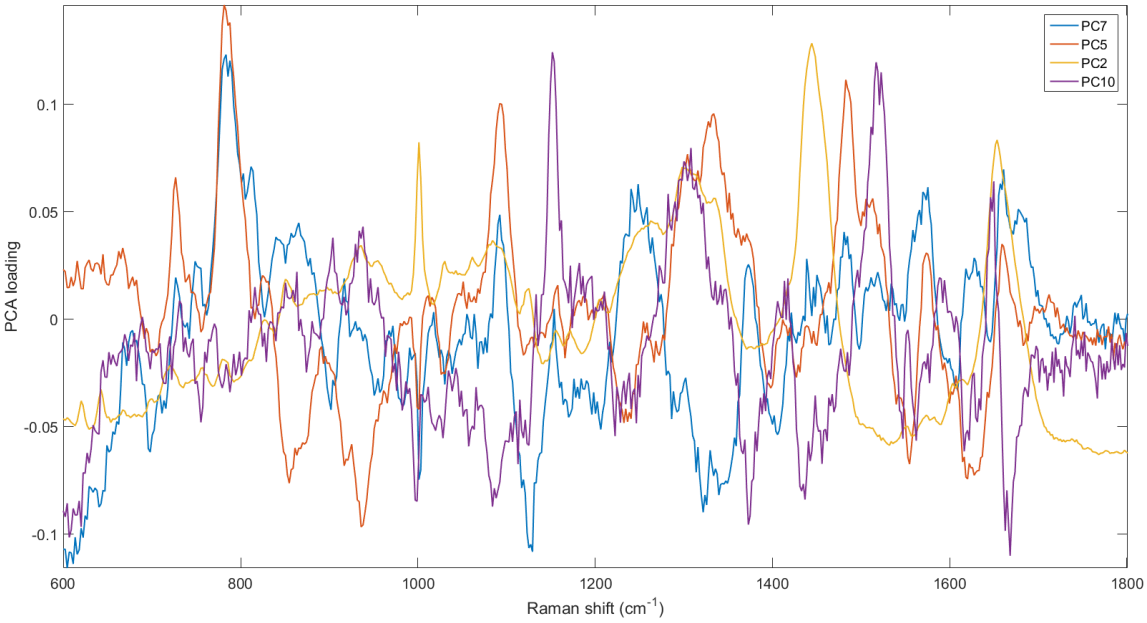

Figure 3.59-PCA loadings highlighting the variability between metastatic and non-cancerous lymph node tissue measurements.

Through comparing the loading peaks to the PCA scores plot (Figure 3.60) it can be deduced that in PC7, peaks more indicative of metastatic data include those assigned as amino acids (640, 698, 1001, 1170 and 1609 cm^{-1}), collagen (1278 and 1322 cm^{-1}) and protein (1406 cm^{-1}). The PC7 loading peaks more indicative of non-cancer include those in the positive direction, such as those assigned as, Nucleic acids and DNA (784, 804, 1092 and 1576 cm^{-1}) and Amide III and I (1249 and 1678 cm^{-1} respectively). The negative peaks in PC5, are more indicative of metastatic samples and are assigned as amino acids (855, 1555 and 1620 cm^{-1}), and the positive peaks, more indicative of non-cancer samples, are assigned as DNA and nucleic acids (727, 781, 1092, 1334 and 1483 cm^{-1}), protein and lipid (1334 and 1483 cm^{-1}), carotenoid (1513 cm^{-1}) and amide I (1659 cm^{-1}). The loading spectrum of PC2 largely consists of positive bands that make up the majority of the peaks belonging to lymph node spectra. The only potentially negative peak was at 1530 cm^{-1} assigned as carotenoid. This might support the observation in the PCA scores plot, where metastatic and non-cancer scores are both more positive.

Table 3.22-Positive and negative bands from the PCA loadings with their tentative assignemnts.

PC7				PC5				PC2				PC10			
-ve		+ve		-ve		+ve		-ve		+ve		-ve		+ve	
640.1	Tyr	783.5	Pyrimidine	854.8	Tyr	726.6	DNA	1530	Carotenoid	619.9	Phe	755.1	Tryp	1152	Protein
697.9	Cystine	811.7	Nucleic acid	935.9	Protein	781.3	Pyrimidine			642.3	Tyr	999	Phe	1308	Lipid, collagen
1001	Phe	865.6	Tyr	1001	Phe	1092	DNA			720	DNA	1028	Phe	1517	Carotenoid
1129	Lipid, protein	1092	DNA	1398	Protein	1334	Carbs, nucleic acid			757.3	Tryp	1057	Lipid	1650	Amide I
1170	Tyr	1249	Amide III	1427	Protein, phospholipid	1371	Tryp			779.1	Pyrimidine	1084	Protein, lipid		
1206	Protein	1576	Nucleic acid	1555	Tryp	1483	DNA			826.	Nucleic acid	1113	Lipid, protein		
1286	Collagen	1628	Amide I	1620	Amino acid	1513	Carotenoid			850.5	Tyr	1222	Amide III		
1322	Collagen	1661	Amide I			1659	Amide I			935.9	Protein	1247	Amide III		
1340	Carbs, Nucleic acid	1678	Amide I							954.9	Hydroxyapatite, cholesterol	1373	Tryp		
1406	Protein									1001	Phe	1437	Protein, phospholipid		
1609	Amino acid									1263	Amide III	1456	Protein, phospholipid		
										1308	Lipid, collagen	1547	Tryp		
										1445	Protein, phospholipid	1562	Nucleic acid		
										1654	Amide I	1620	Amino acid		
												1668	Amide I		

PLSDA was then carried out on the dataset to evaluate the capabilities of the supervised method to classify metastatic lymph nodes from reactive lymph nodes, from the three anatomical locations. Figure 3.60 shows the supervised clustering of the two groups of the dataset, although visually not overly distinct

from each other, cross-validation resulted in a sensitivity of 71% and specificity of 89% for the identification of metastasis in lymph nodes (Table 3.23).

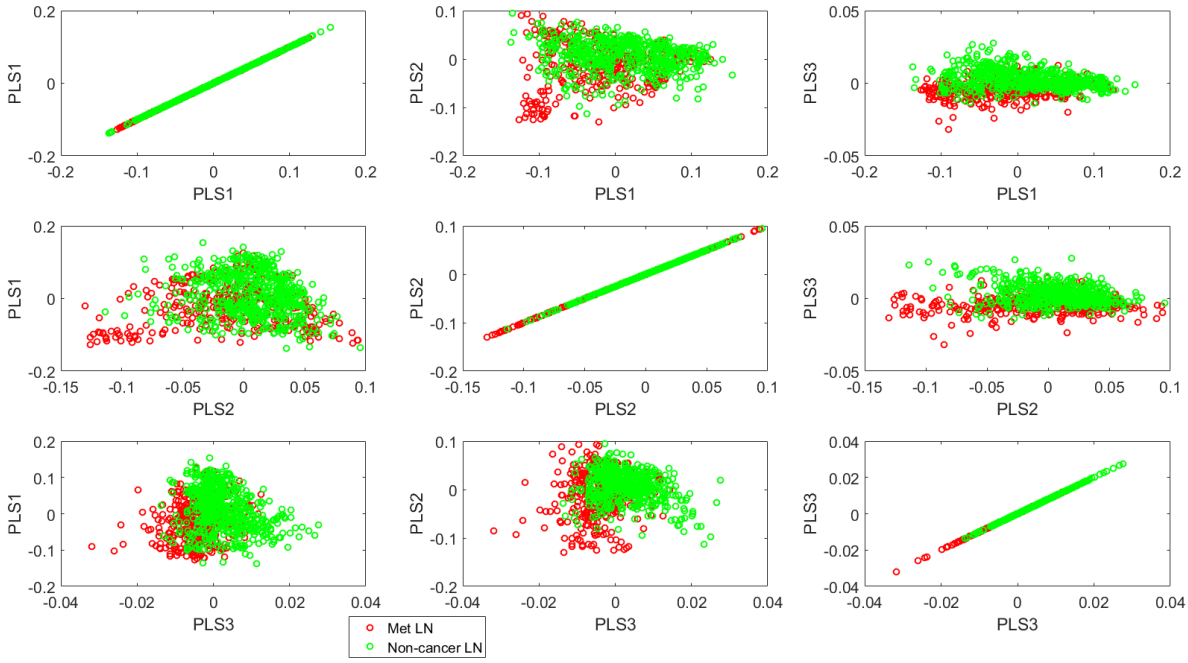

Figure 3.60-PLSDA scores plot for the first three components, showing the distribution of metastatic (red) and non-cancer (green) LN measurements.

Table 3.23- Confusion matrix for the classification of metastatic LN from non-cancer LN.

		True Group	
		Metastatic	**Non-cancer**
Predicted group	**Metastatic**	12	3
	Non-cancer	5	24
Total LN		17	27
Sensitivity		0.7059	0.8889
Specificity		0.8889	0.7059

The PLS scores plot for component 1 and 2 was plotted, Figure 3.61, and labelled with the anatomical location of the LN as well as the pathology, to assess if clustering between lymph nodes of the same location, but different pathology, was occurring. From the plot it does not seem like this is taking place, as the LN data from the different locations are intermingled with each other.

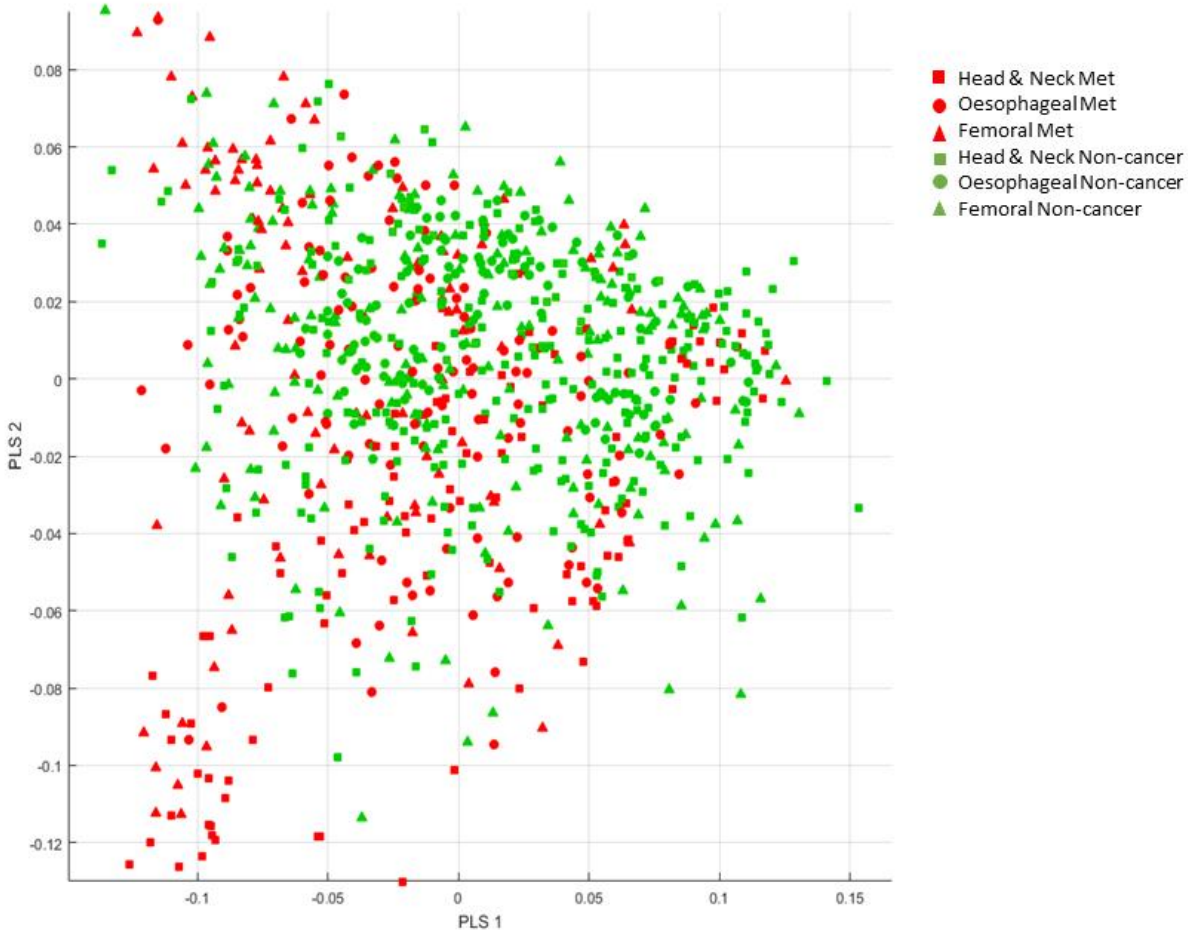

Figure 3.61- PLSDA scores plot of component 1 and 2 and labelled by both pathology and location.

The VIP plot displayed in Figure 3.62 shows the most important spectral wavenumbers for distinguishing metastatic LN from non-cancer nodes and Table 3.24 lists those selected 100% of the time for cross validation with their corresponding tentative assignments. Comparing this to the PCA component loadings, the peaks primarily indicative of metastatic LN are those assigned as amino acids (645, 695, 1001, 1549, 1620 cm^{-1}), lipid and protein (1129 cm^{-1}), collage (1322 cm^{-1}) and carbohydrate (1340 cm^{-1}). Those mostly indicative of non-cancer nodes consist of those assigned as DNA and amino acids (784, 1094 and 1579 cm^{-1}), carbohydrate (1334 cm^{-1}), protein and phospholipid (1483 cm^{-1}) and carotenoid (1513 cm^{-1}).

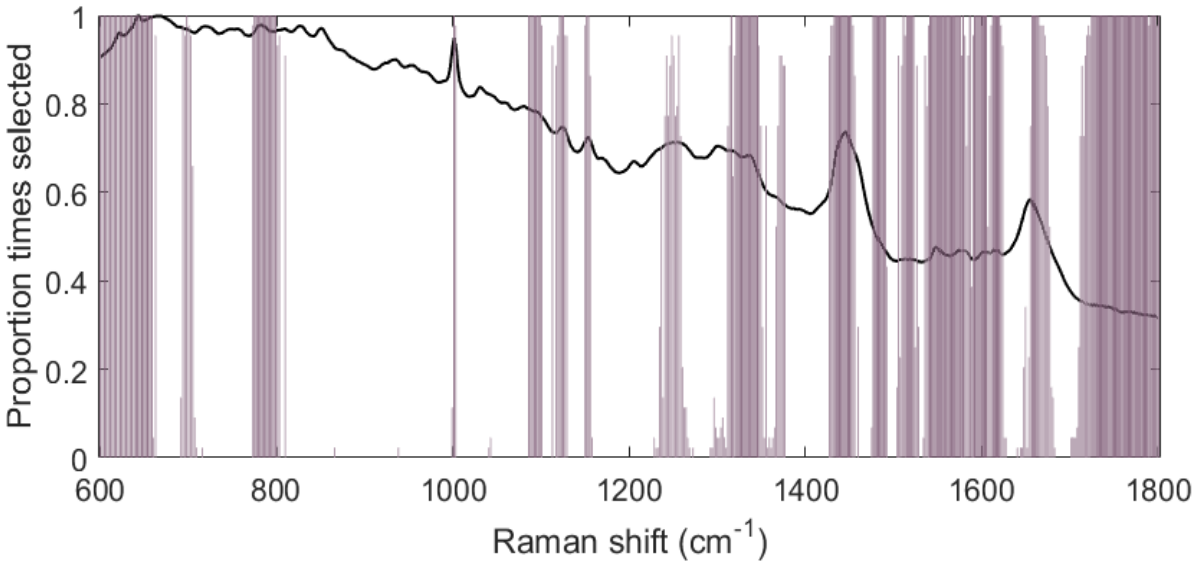

Figure 3.62- VIP plot showing the proportion of times specific Raman shifts were selected to discriminate between the groups of the dataset in PLSDA cross-validation.

Table 3.24- Spectral regions from VIP selected 100% of times for cross-validation of PLSDA model.

Bands selected 100% of time for classification	Tentative assignment
599.7-657.9	Phe, Tyr
697.9	Glycine
774.7	Pyrimidine base
1001	Phe
1086-1099	Protein, lipid, DNA
1121-1125	Lipid, protein
1152-1154	Protein
1316	Collagen
1322-1345	DNA, nucleic acid
1433-1452	Protein, phospholipid
1477-1490	Protein, phospholipid
1511	Carotenoid
1515-1523	Carotenoid
1542-1575	Tryp, Nucleic acid
1587	Nucleic acid
1592-1605	Amino acid
1613-1620	Amino acid
1657-1665	Amide I
1725-1788	Lipid
1792-1797	-
1801	-

3.2 Discussion

Lymph nodes are primarily located within and attached to adipose tissue.[231] Adipose tissue is loose connective tissue that stores energy in the form of lipids. Lipids, along with proteins and carbohydrates, are the main constituents of the human body, hence is stored in many cells throughout different tissue types in different quantities. Lymph fluid contains a high concentration of lipids and fat deposition often occurs in the lymphatic structures. Also, the leakage of lymphatic fluid from its vessels appears to promote the accumulation of fat.[232] High concentrations of adipocytes surrounding lymph nodes has been found to be related to chronic lymphadenopathy, inflammation, of peripheral lymph nodes.[233] A high fat content was observed in the oesophageal lymph nodes and all were typically surrounded by adipose tissue, shown through its spectrum in Figure 3.27. As well as the lymph nodes possessing a naturally higher lipid cell content than other tissue, some sections exhibited pools of fat. Suggesting that in some sections, surrounding adipose might have been dragged across the tissue during the sectioning process. This creates a challenge in obtaining biochemical signal from the lymph node tissue that is not overwhelmed by intense lipid peaks. This especially causes problems when mapping areas of the tissue sample.

The carbon particles that pigment the lymph node sections produce challenges for spectral acquisition. Anthracosis of lymph nodes is caused as a result of the inhalation of carbon particles into the lungs followed by their removal via macrophages in the lymphatic system and transportation to the lymph nodes.[234] Anthracosis is a form of pneumoconiosis, a benign disease with unknown clinical implications, associated with tuberculosis, infection, smoking, air pollution and biomass.[235] The histiocytes that engulf the carbon pigments situate in the medulla of the lymph node. Metastatic cancers also infiltrate into the lymph node via the afferent lymphatics. The tumour cells then spread throughout the subcapsular and medullary sinuses and eventually to the whole node. Thus, tumour cells are initially seen in the subcapsular and medulla regions. However, these carbon particles cause difficulty when measuring the medulla of the nodes, where the metastatic cancer often presents, far more difficult and limits the number of areas where viable spectra can be measured.

Severe anthracosis has been reported to reduce the lymph nodes ability to efficiently trap antigens which can eventually pass straight through the node and can also cause false positive diagnosis using PET scans during the staging of lung cancer.[236] Lymph nodes from below the diaphragm have lower carbon particle content than nodes above due to reduced drainage from the lungs, hence spectral measurements are more representative of biochemical information of the lymph nodes and do not result in intense spectral backgrounds as a result of carbon particles. Therefore the lymph nodes from these locations were acquired when possible for spectroscopic analysis of oesophageal nodes in this study, to enable better measurements of nodal tissue.

The challenges of fat deposits, carbon pigmentation and tissue scoring of the oesophageal tissue led to the need for the primary acquisition of point spectra in this part of the study as opposed to tissue mapping. This method of spectral acquisition resulted in lower spectral acquisitions but ensured spectra weren't affected by fat or carbon contributions as the quality of each were visually checked. Point collection also more closely resembles how measurements would be taken *in vivo*. Only a few sections from the sample set produced sufficient Raman maps to permit evaluation of digital staining shown in Figure 3.32.

Primary cancers of the lymph node are diagnosed based on the lymphocyte they originate from, either B-cell or T-cell, and the location in the lymph node. Therefore the ability to distinguish between these two cell types is important for diagnosis of the pathology of a tissue sample. The diagnosis may also influence the treatment and thus an accurate diagnosis is vital for the understanding of the treatment available for a patient and their prognosis. The Raman map, Figure 3.32, shows that the T-cells and B-cells in one sample can be identified and distinguished from one another using K-means clustering, demonstrating that the two cells must be spectrally different from one another based on the biochemicals detected. Romeo and Diem[237] were able to discriminate the regions within the lymph node based on clustering of the infrared spectral map.

Many studies looking at Raman spectroscopy as a diagnostic technique have found that the intensity of the lipid bands are lower and the protein bands

higher in cancerous tissue compared to non-cancerous tissue.[199] This is attributed to the replacement of tissue lipids with rapidly dividing tumour cells, causing an increase in protein and a decrease in lipid concentrations. The measurements from this study show the opposite however, with an increase in lipid and a decrease in protein in metastatic AC tissue. Adipocytes are a major constituent in the tumour microenvironment and a positive relationship between obesity and cancer, especially oesophageal cancers, has been observed by many epidemiological studies.[238] Renehan *et al*.[239] found that oesophageal adenocarcinoma is the highest cancer in males and the third highest in females associated with obesity. This may provide an explanation for the increased lipid peak intensities in measurements from nodes with metastatic AC from oesophageal tumours compared to non-cancer lymph nodes. Furthermore, the adipose deposits surrounding lymph nodes respond directly to stimulation from the immune system to provide a store of energy when needed. The lymphatic fluid also has a strong pro-adipogenic effect and adipose progenitor cells can be reprogrammed into lymphoid organiser cell to contribute to stroma.[240]

Lloyd *et al*.[58] note that metastatic invasion can reduce the number of nucleic acids due to a higher cytoplasm to nucleus ratio in epithelial cells compared to lymphocyte cells and note that lymphocytes have been shown to display strong nucleic acid signals[241] which may offer an explanation for the decrease in DNA peak intensity in the metastatic than the non-cancer sample measurements. However, this decrease in DNA content does disagree with the study by Isabelle *et al*.[208] where they found lipid concentration and carbohydrate at lower signals in benign tissue measurements and DNA higher.

During tumour progression the secretion of type VI collagen is upregulated.[242] Amide III is a component of collagen within tissues. Higher collagen peak intensities in the cancerous lymph nodes may be attributed to the desmoplastic response which is induced by malignancy. Growth factors that are secreted by cancer cells activate fibroblasts to synthesise collagen.[243] The apparent increase in collagen that was observed in the metastatic lymph node measurements of this study is also frequently seen in breast tissue with malignancy and is consistent with other Raman spectroscopy studies. This increase is attributed to fibrosis associated with lesion formation.[244] Denser

breast tissue is associated with increased collagen content as a result of breast cancer and also aids mammography in identifying tumours. Additionally, women with dense breast tissue are found to have an elevated chance of developing breast cancer at some point during their lives.[245] Therefore, whether the collagen increase is caused by cancer, the collagen increase aids the progression of cancer, or they are both positively related, would be of interest in better understanding the development of cancer and its diagnosis.

The health benefits of carotenoids have been reported for preventing a variety of diseases, such as cancers and cardiovascular diseases.[246] A decrease in carotenoid levels in the metastatic samples observed in this study was also reported by Bakker-Schut *et al.*[247] *They* found carotenoid concentrations in lymphocyte tissue at decreased levels in cancer samples than compared to healthy. However, they also saw carotenoid levels decrease with age and found that while there was a significant difference between all their 'healthy' and 'lung cancer' samples, there was not a significant difference between the 'healthy old' and 'lung cancer'. Therefore, it would only be reliable to take this into consideration and compare the carotenoid levels within age groups. When producing classification models for clinical diagnosis, it could be more reliable to have a number of different models trained based on age range that a blind sample could be tested against depending on the patient demographics. Hata *et al*.[248] also reported carotenoid signals at a decreased level in neoplastic and precancerous human skin when compared to healthy tissue.

The glycine assigned peak seen in the metastatic spectrum may be the result of elevated levels of glycine when a tissue becomes necrotic. Sitter *et al.* (2007)[249] also observed an increase in glycine levels in tissue samples from large breast tumours using MRI and attributed this increase to a connection with the synthesis of collagen, which is also found in higher concentrations in cancerous tissue.

PLSDA, an unsupervised method, was used immediately without a prior unsupervised technique. This was done as biochemical differences between cancer and non-cancer samples are already known to exist and have previously been detected using Raman spectroscopy.

The PLSDA scores plot for the oesophageal lymph node tissue, Figure 3.36, shows the grouping of the data based on the top three components and shows slight separation of the groups. Cross validation of the PLSDA model produced a sensitivity of 50% and specificity of 83% for the identification of oesophageal lymph node samples with adenocarcinoma metastases. The weights further show that there are stronger DNA and carotenoid peaks associated with non-cancer measurements, and stronger collagen and lipid peaks associated with metastatic adenocarcinoma in oesophageal nodes.

A positive peak present at 1296 cm^{-1}, attributed to CH_2 vibrations, exhibited in the weights of PLS component 2 indicates more intense presence in non-cancerous tissue spectra than those from metastatic cancer. Whilst other lipid signal produced also from CH_2 bonds as well as C=C and C=O bonds are as a whole more intense in metastatic tissue measurements. Malini *et al.*[217] also reported the stronger presence of this peak at 1301 cm^{-1} when measuring the signal from inflamed oral tissue samples than when compared to malignant and premalignant tissue. The non-cancerous tissue samples used in this study were often diagnosed as minimally reactive where, like inflammation, they are responding to infection. This may provide an explanation for the peak's stronger presence in non-cancerous tissue, whilst lipid molecules as a whole are still more abundant in metastatic tissue, and why it may aid discrimination between the pathological groups.

A hydroxyproline peak at 865 cm^{-1} was identified as important by VIP for identifying metastatic AC in lymph node tissue. Hydroxyproline is an amino acid unique to collagen which was also found to be at elevated levels in met AC tissue, observed by an increase in the associated Raman bands. Studies have found that patients with bone metastases had elevated levels of excreted hydroxyproline.[250,251] Li *et al.*[252] observed a higher level of hydroxyproline in blood samples of patients with ovarian cancer than those without the disease or other benign conditions. They found that the level of hydroxproline was not affected by age, but pregnant women can have levels higher than usual. The tryptophan assigned peak is observed at lower levels in met AC tissue seen in the PLS weights in component 1 at 1551 cm^{-1} and component 3 at 1559 cm^{-1} in the

positive direction. Krishna *et al.*[253] also observed a decrease in these peaks in FTIR spectra in malignant and benign ovarian tissue.

VIP identified the pyrimidine bases of DNA, phenylalanine, carotenoid, lipid, collagen, nucleic acids, amino acids and amide I regions as important for classification. Amide I appears in both the positive and negative weights and therefore may suggest the presence of a peak shift between the cancer and non-cancer group as oppose to a difference in its intensity between the two. In the metastatic cancer tissue it appears to be detected at a slightly higher wavenumber than in the non-cancer. This could be suggestive of a change in microenvironment, resulting from the infiltration or development of cancer, which may lead to slightly altered Raman scattering.

Visually a notable difference between the femoral and oesophageal lymph node measurements is much less lipid contribution in the femoral LN samples.

The PLSDA model was cross-validated resulting in a sensitivity and specificity of 83% and 75% for the identification of femoral lymph nodes with squamous cell carcinoma metastases. Like the oesophageal lymph node measurements, DNA and carotenoid peaks were observed to be more characteristic of non-cancer LN measurements than cancer. Lipid, phenylalanine and collagen peaks were similarly found to be more indicative of the metastatic LN measurements. Amide I was again observed in both the negative and positive weights and might further indicate a difference in Raman shift rather than quantities in the samples. The amide I appears to be present in the metastatic samples at a decreased Raman shift to the non-cancer samples, unlike the oesophageal metastatic samples. The assigned amide III peak appears more intense in the non-cancer samples, suggesting its presence at a decreased level in femoral LN samples that are metastatic.

A hydroxyapatite/ cholesterol peak is present in the negative weights and more indicative of the metastatic measurements. Sathyavathi *et al*.[254] observed the presence of hydroxyapatite, at 960 cm^{-1}, and found it to be related to microcalcifications in both malignant and benign tissue. This suggests that both cancer and non-cancer tissue may exhibit a more intense hydroxyapatite peak in

relation to the presence of calcifications. However, a strong indication of a malignancy is the presence of microcalcifications and thus provides an explanation for why in this study stronger hydroxyapatite peaks are indicative of the metastatic samples.

Again, the lipid contributions appear much less in the head and neck LN measurements than those of the oesophageal LN. The initial visual differences coincide with the other two lymph node areas where DNA and nucleic acid peaks are less intense in the metastatic measurements and lipid is stronger.

The scores plot of PLSDA was able to visually separate the two groups of the dataset better than either the femoral LN or the oesophageal LN data. This better separation was reflected in the cross-validation of the model where a sensitivity of 80% and specificity of 92% was calculated for the identification of head and neck lymph nodes with SCC metastases.

The weights from the PLSDA, highlighting peaks in the spectra contributing to variation between the groups, include peaks more representative of non-cancer, assigned as DNA, carotenoid, nucleic and amino acids and amide III. Peaks more representative of metastasis primarily include those assigned as lipid, phenylalanine, collagen and hydroxyapatite. These results support those observed in the previous oesophageal and femoral LN PLSDA models.

The VIP also highlighted similar areas of the spectrum as important for classification as the oesophageal and femoral LN dataset. Areas of the spectrum that were identified that were also in the VIP of the femoral and oesophageal LN consisted of the areas around the following assigned peaks: the pyrimidine base of DNA (784 cm^{-1}), phenylalanine (1001 cm^{-1}), collagen (1302 cm^{-1}), amide III (1338 cm^{-1}), protein/phospholipid (1444 cm^{-1}) and amide I (1656 cm^{-1}). These areas of the spectrum could be where the general variation between metastasis and non-cancer lymph nodes occur, and also where the similarities between the three non-cancer lymph node areas exist.

The PLSDA of the mapped data resulted in a lot of similarity to the point measurement data, and the cross-validation carried out achieved the exact same

sensitivity and specificity. This result shows that a high number of measurements, that causes a long overall acquisition, is not necessarily needed for improved classification and that just a few measurements are sufficient to get an appropriate spectral representation of the sample. This result also gives further support to real time diagnosis and the feasibility of only taking point measurements *in vivo*. The main difference between the point and mapped datasets is the decrease in noise in the PLSDA component weights in the mapped dataset, although the main peaks are the same. A difference in the VIP is seen with the mapped data which includes the carotenoid peak (1154 cm^{-1}) and lipid (1723-1801 cm^{-1}) as important for classification, and lacks the wavenumber regions of the phenylalanine and amide III assigned peaks.

It is clear from the k-mean cluster image, Figure 3.53, that 'digital staining' could be a possibility for samples that have metastatic cells present, to remove the need and expense of physically staining the samples. Diem *et al*.[255] shows the ability for "molecular pathology" using vibrational micro-spectroscopy for the identification of metastases in lymph nodes, early morphological changes in oral and cervical tissue and distinguishing between SCC and AC metastases through the recognition of different cell types. Nallala *et al*.[256] also report their ability to use FTIR for spectral histopathology to aid standard approaches through analysis of the same tissue section using digital dewaxing[257]. Although the collection of spectral maps is a much more lengthy process than taking point acquisitions, this method could increase sample throughput compared to the current methods that can take days to produce stained sections. Fernandez *et al*.[258] suggest three aspects to be considered for the implementation of digital staining: rapid and reliable data acquisition, the development of robust algorithms and testing its validity. The clear margin between the non-cancer lymphocytes and the SCC cells also supports the potential for tumour margin delineation which has the ability to aid surgery in real-time. Diem *et al*.[255] state that brain metastases have well defined tumour margins and reports the ability to use spectroscopy for aiding tumour removal in the brain, where only removing necessary tissue is essential for preserving brain function of the patient after the operation. The figure also shows the heterogeneity of the samples and the difficulty that can be experienced in measuring similar regions for each sample.

Keratinocytes are often present in the tissue of SCC metastases in lymph nodes and its presence often can confirm diagnosis. Kim *et al.*[259] observed that changes to vitamin A sensitivity in malignant cells causes a decreased differentiation of these cells and thus keratin synthesis. They also noted that there was an alteration between the synthesis of type I and type II keratin between healthy and malignant cells. The three different clusters, seen by the three colours, in the k-means image could perhaps represent different keratinocyte types and a better understanding may be important in recognising SCC cells.

FTIR was conducted to compare it to Raman spectroscopy and enable analysis of the two techniques using the same samples and region of interest (ROI). For *in vivo* analysis Raman spectra are not affected by water contributions where FTIR spectra are. However, if FTIR is a much better technique than Raman for classification when used *ex vivo*, it could be more important to focus on developing the appropriate pre-processing techniques that help combat the spectral water features, to enable *in vivo* use of FTIR. However, this research found that FTIR actually had a lower sensitivity and specificity than Raman spectroscopy for classifying lymph nodes with SCC metastases from non-cancer samples. The FTIR spectra are in agreement with the Raman, regarding less intense nucleic acid assigned peaks in the averaged metastatic data compared to the non-cancer.

PCA of the combined sample set of oesophageal, femoral and head and neck lymph node data enabled the analysis of the unsupervised separation that occurs. The PCA scores plot did not give any indication of data groupings other than pathologically. PCA recognised many of the same peaks as the previous PLSDA models did, as variation between the metastatic cancer and non-cancer measurements. These include collagen, DNA, amide I and III and lipid assigned peaks.

The PLSDA scores plot did not show separation of the data based on anatomical location or metastatic type. Cross-validation of this model produced a sensitivity of 71% and specificity of 89% for identifying metastatic measurements from non-cancer. This shows that potentially metastatic lymph node cancers can be differentiated from non-cancer lymph nodes when in a combined model.

Although, the accuracy has decreased for that of the head and neck where the sensitivity and specificity was much better. Therefore, for these samples it might be more beneficial to have a separate classification model. The VIP showed the majority of spectral variation was in the DNA and nucleic acid, collagen, lipid and carotenoid regions.

3.3 Conclusion

Overall this study has demonstrated the use of Raman spectroscopy for analysing lymph node tissue from three different anatomical locations.

It has been observed that lymph nodes from the oesophagus can contain anthracosis that cause difficulty when measuring certain areas within the tissue, and can even result in overwhelmed signal to the spectrum. The lymph nodes from this location also appear to have a larger amount of adipose tissue surrounding and embedding it, observed through strong lipid peaks that can also affect resulting Raman spectra. Therefore, these lymph nodes are difficult to measure however point spectra were produced with avoidance of areas affected by anthracosis and adipose tissue. The difficulty to measure these lymph nodes were perhaps reflected by poor sensitivities and specificities of just 50% and 83% for the identification of LN with adenocarcinoma metastases from non-cancer oesophageal nodes.

Femoral lymph node measurements were far easier to carry out and no anthracosis or intense lipid peaks were observed. A sensitivity of 83% and specificity of 75% was calculated for identifying squamous cell carcinoma from non-cancer femoral nodes.

Head and neck nodes were also easier to acquire spectra from and no anthracosis or strong lipid peaks were observed. The resulting sensitivity and specificity from both the point measurements and spectral maps were 80% and 92% for the identification of LN with squamous cell carcinoma metastases from non-cancer head and neck nodes. The ability to classify at the same level from just 20 spectra per sample, supports the viability for real-time *in vivo*

measurements that rely on just a few measurements to produce representative data of a sample. FTIR spectroscopy did not prove to be a better technique for disease classification in this study and therefore a focus on Raman could be continued in subsequent experiments. Raman spectroscopy has also been shown to produce good spectral maps that can be coupled with clustering techniques in order to digitally differentiate different regions with the lymph nodes and even identify metastatic regions and their border. This result indicates the potential for Raman digital staining of lymph node metastases to reduce time and expense of diagnosis. The clear cancer borders identified through k-means clustering could be a promising indication of its surgical capabilities. A larger study based on the analysis of Raman maps would be needed to fully assess its clinical capabilities as a histological tool for tissue sections.

The ability to combine lymph node regions for metastatic detection as a whole has also been demonstrated. But whilst it helps to improve oesophageal classification it would be less beneficial for femoral and head and neck nodes that had better sensitivities and specificities in their independent models. Also, the differences between the lymph nodes from the different anatomical regions is not known and how this would impact on the reliability of the data.

Chapter 4: Raman Analysis of Lymph Node Regions

4.1 Results

Initially measurements from the non-cancer lymph nodes of the head and neck, oesophageal and femoral regions, were analysed using PCA to assess if there were any biochemical differences between the anatomical regions that had led to a substantial difference in the Raman signal detected. If all lymph nodes are spectrally the same then their analysis can be combined to create a larger diagnostic model. If however spectral irregularities are observed between nodes from different areas, it may indicate that more specific diagnostic models, of not just the tissue type but the anatomical region also, are required for Raman spectroscopy to better indicate disease.

Figure 4.63 shows the spectra of the mean head and neck, oesophageal and femoral lymph node data. Differences between the spectra appear primarily due to variation in spectral intensity rather than the shifts in the peak positions.

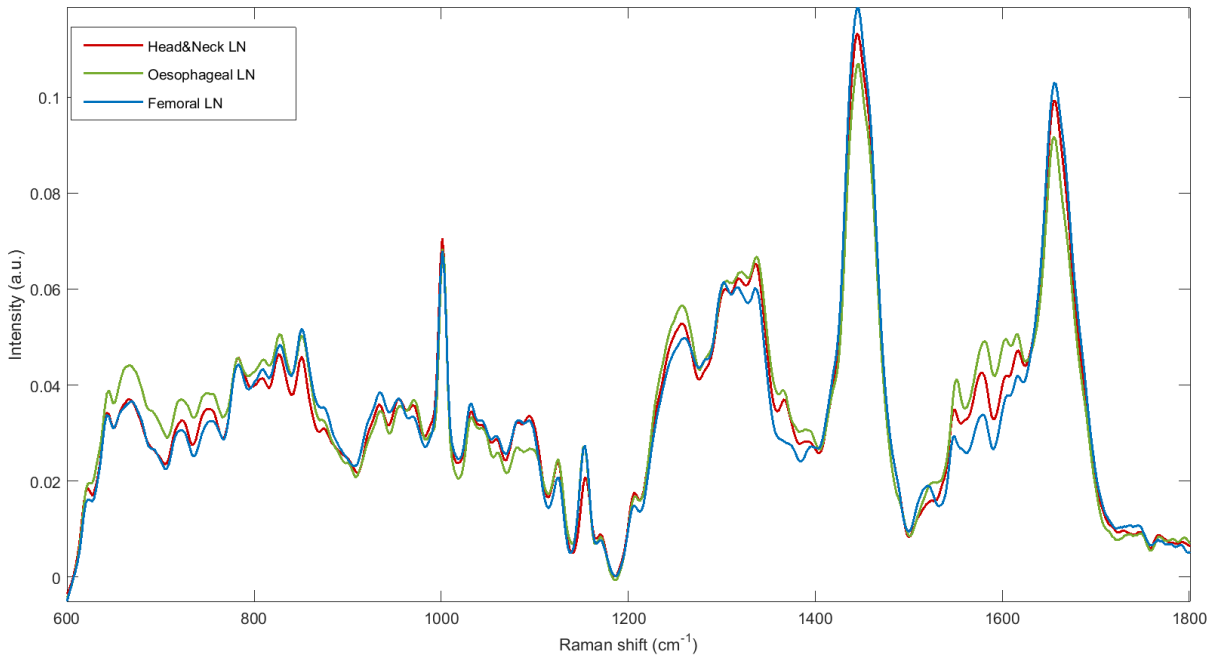

Figure 4.63- Spectra averaged from 260 head & neck (red), 120 oesophageal (green) and 160 femoral (blue) lymph node measurements.

Figure 4.64 shows the PCA score plots from the top four components for variation between the groups with at least a 99% confidence calculated using ANOVA. The scores from the three groups show a large amount of overlap with each other. Clustering is present however, and the scores from the groups appear to congregate together at similar areas on the component axes. The green oesophageal cluster appears the most visibly distinct from the other groups. Principal component six is responsible for the most variance between the groups. In this component the oesophageal node scores are primarily grouped on the negative axis whilst the head and neck and femoral node scores are most frequently positive. The scores for PC3 are mainly negative for femoral node data and positive for oesophageal node data, whilst the head and neck node scores are equally distributed across the component. In PC 9, it appears to separate head and neck LN data, distributed more across the negative axis, from data of the other two LN regions, which are generally more positive. In PC8 the data from all three LN regions are quite dispersed across both the negative and positive axis and therefore it is difficult to see a separation between groups.

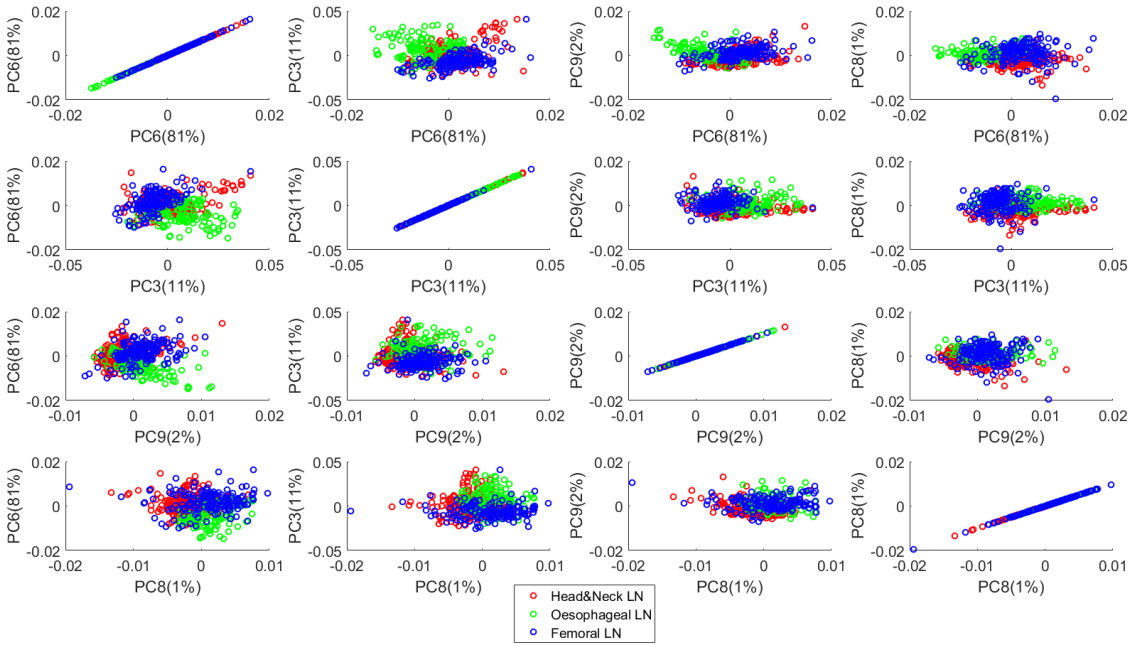

Figure 4.64- PCA scores plots of lymph node measurements from the head and neck (red) oesophageal (green) and femoral (blue) regions.

The PCA loadings, displayed in Figure 4.65, show the peaks in the spectra that are responsible for the variation between the clusters in the PCA scores plot.

The loadings appear to consist of many peaks across the whole of the 600-1800 cm^{-1} wavenumber region. The main peaks from the loadings are displayed in Table 4.25, separated by their either positive or negative direction to enable identification of the possible biochemicals that are more frequent in a particular lymph node group.

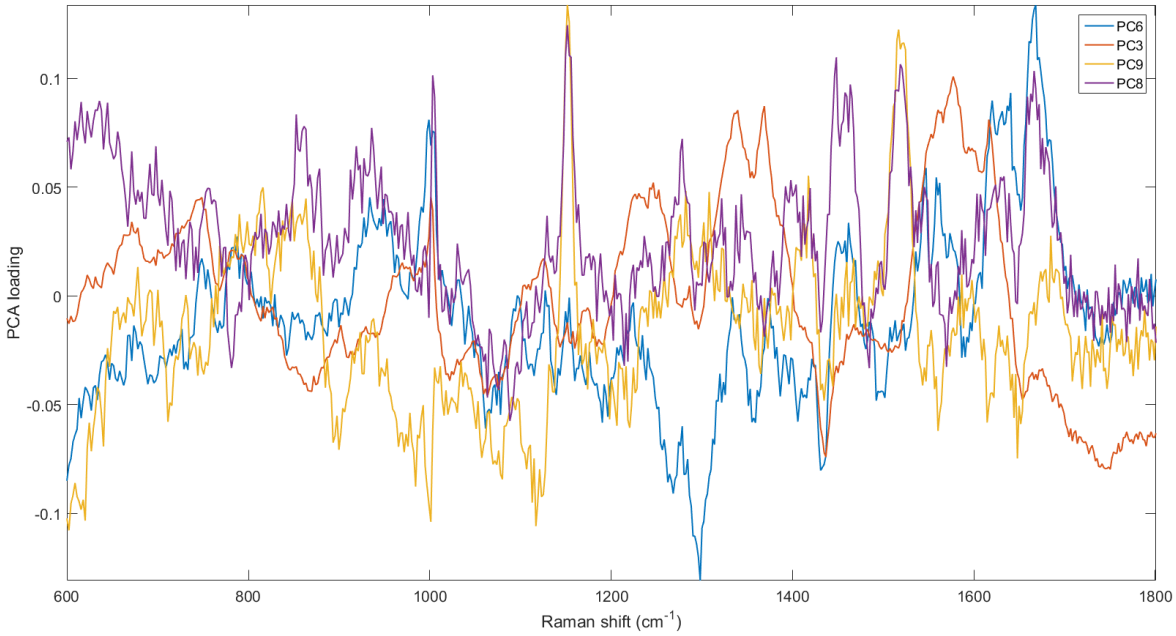

Figure 4.65-PCA loadings of the top four components for discriminating head & neck, femoral and oesophageal LN data.

In PC 6, positive bands include those assigned to amide I (1631, 1641 cm^{-1}) and amino acids (999, 1003, 1559, 1562, 1620 cm^{-1}) whilst the predominant negative band is assigned to amide III (1268 cm^{-1}). According to the distribution of data in the scores plot, amide III peaks appear to be more frequent in spectra belonging to oesophageal LN whilst amide I and amino acids, such as phenylalanine and tryptophan, bands are more frequent in those from head and neck and femoral LN. Also, among the negative bands is a lipid/collagen (1298 cm^{-1}) peak, suggesting a higher level in oesophageal LN than when compared to the other LN locations. In PC 3 the negative bands consist of peaks assigned to lipids (1437, 1745 cm^{-1}), whilst the positive bands include those assigned to amide III (1231, 1247 cm^{-1}), carbohydrate (1340 cm^{-1}) and nucleic acids (1561 cm^{-1}). In this component the scores plot suggests oesophageal LN data are primarily on the positive axis whilst femoral LN measurements are on the negative

axis. Therefore, amide III appears to be more indicative of oesophageal origin, along with carbohydrates and nucleic acids, than femoral lymph node data. Negative peaks, more indicative of femoral LN, constitute the main lipid peaks. PC9 appears to separate head and neck lymph node data, which congregate mainly on the negative axis, from the other two lymph node regions. Negative peaks from this component include amino acids (619, 640, 750, 1001, 1561 and 1615 cm^{-1}) and amide I (1648 cm^{-1}) whilst the positive peaks are those of amide III (1282 cm^{-1}), lipids (1308 and 1418 cm^{-1}) and carotenoid (1517 cm^{-1}). The bands in PC8 are primarily peaks assigned as amino acids and proteins, however it is not obvious which clusters positive and which are negative on the PCA scores plot.

Table 4.25-Positive and negative bands from the PCA weights in Figure 3.

PC6				PC3				PC9				PC8			
-ve Assignments		+ve Assignments		-ve Assignments		+ve Assignments		-ve Assignments		+ve Assignments		-ve Assignments		+ve Assignments	
695.7	Cystine	933	Protein backbone	869.8	Tyrosine	748.5	DNA	619	Phe	816	Phosphodiester	781	Pyrimidine	622.2	Phe
1061	Lipid	999	Phe	1437	Proteins phospholipids	1001	Phe	640	Tyr	1152	Proteins	1088	Protein, lipid	644.6	Tyr
1268	Amide III	1003	Phe	1745	Lipid	1231	Amide III	711.2	DNA	1282	Amide III			697.9	Cystine
1298	Lipid/collagen	1547	Proline			1247	Amide III	750	Tryp	1308	Lipid/collagen			755.1	Tryp
1359	Protein/lipid	1559	Tryp			1340	Amide III/ Carbs/ Nucleic acid	899.8	Adenine	1418	Protein/phospholipid			852.7	Tyr
1431	Lipid	1562	Tryp			1369	Tryp	1001	Phe	1517	Carotenoid			935.9	Protein
		1620	Amide I/ Amino acid			1561	Nucleic acid	1080	Phospholipid					1003	Phe ring
		1631	Amide I			1577	Nucleic acids	1123	Protein, lipid					1152	Protein
		1641	Amide I			1598	Amino acid (phe, tryp)	1561	Tryp					1278	Protein
		1667	Amide I			1617	Amino acid (tyr, tryp)	1615	Amide I/ Amino acid (tyr, tryp)					1448	Protein, phospholipid

								1648	Amide I					1458	Protein phospholipid
														1464	Lipid
														1519	Carotenoid
														1667	Amide I

The clustering between the groups in PCA suggested that biochemical differences existed between the lymph nodes from different anatomical regions potentially allowing them to be identified from one another. Lloyd *et al*. found that PLSDA should produce either better or the same classification accuracies as PCA-LDA as PLSDA chooses the number of components at its optimum performance. Therefore, PLSDA, a supervised technique, was applied in order to assess if the differences were substantial enough to enable classification into specific groups. The PLSDA scores plot in Figure 4.66 show the data clustering of the three groups. A leave-one-sample-out cross-validation was carried out and produced sensitivities between 63-92% and specificities between 71-100% for classification of the measured lymph node regions, shown in Table 4.26. Model performance should increase with a greater number of samples. Cross validation, where one sample is removed from the training set and blind tested, ensures that the model's performance when a sample is removed is accounted for.

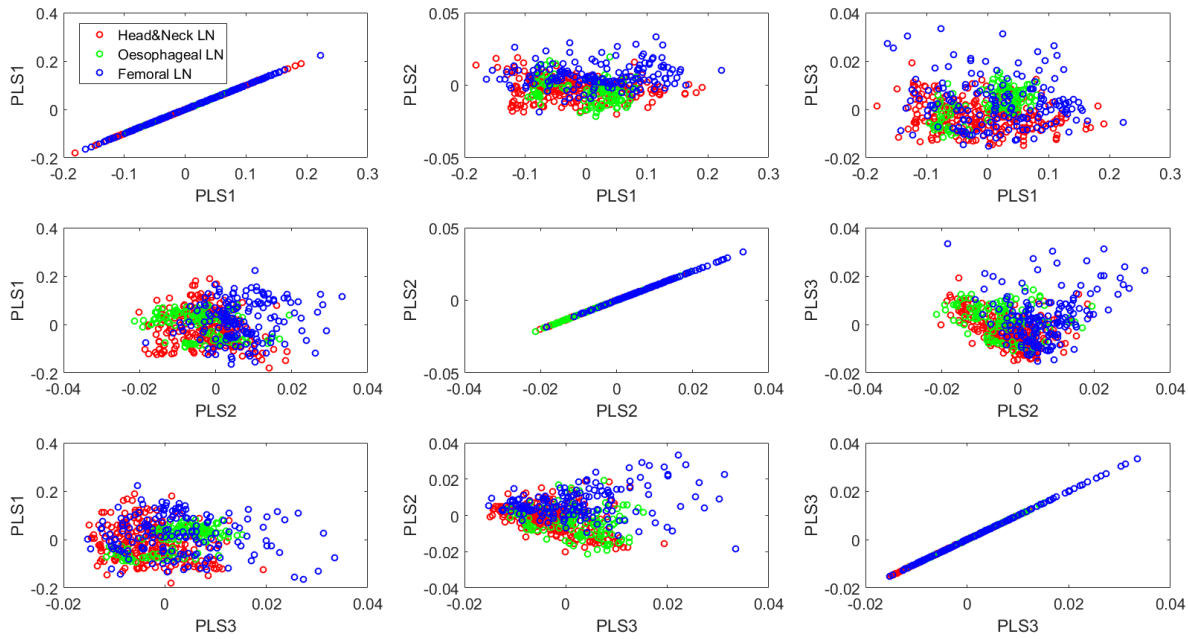

Figure 4.66- PLSDA scores plot showing clustering of data from the different lymph node regions.

Table 4.26- Confusion matrix for the classification of head and neck, oesophageal and femoral LN data.

		True group		
		Head & Neck	**Oesophageal**	**Femoral**
Predicted group	**Head & Neck**	12	1	3
	Oesophageal	0	5	0
	Femoral	1	0	5
Total LN samples		13	6	8
Sensitivity		0.9231	0.8333	0.6250
Specificity		0.7143	1.0000	0.9474

VIP was carried out to assess which bands in the spectrum are important for classification through quantifying the proportion of times those wavenumbers were selected in the cross validation of the PLSDA model. The VIP plots in Figure 4.67 highlight the spectral regions that have been used in cross validation for delineating between the groups and the proportion of times they have been selected. Several bands across the spectrum were selected for cross validation therefore, Table 4.27 displays only the wavenumbers of the bands selected 100% of the time for classification.

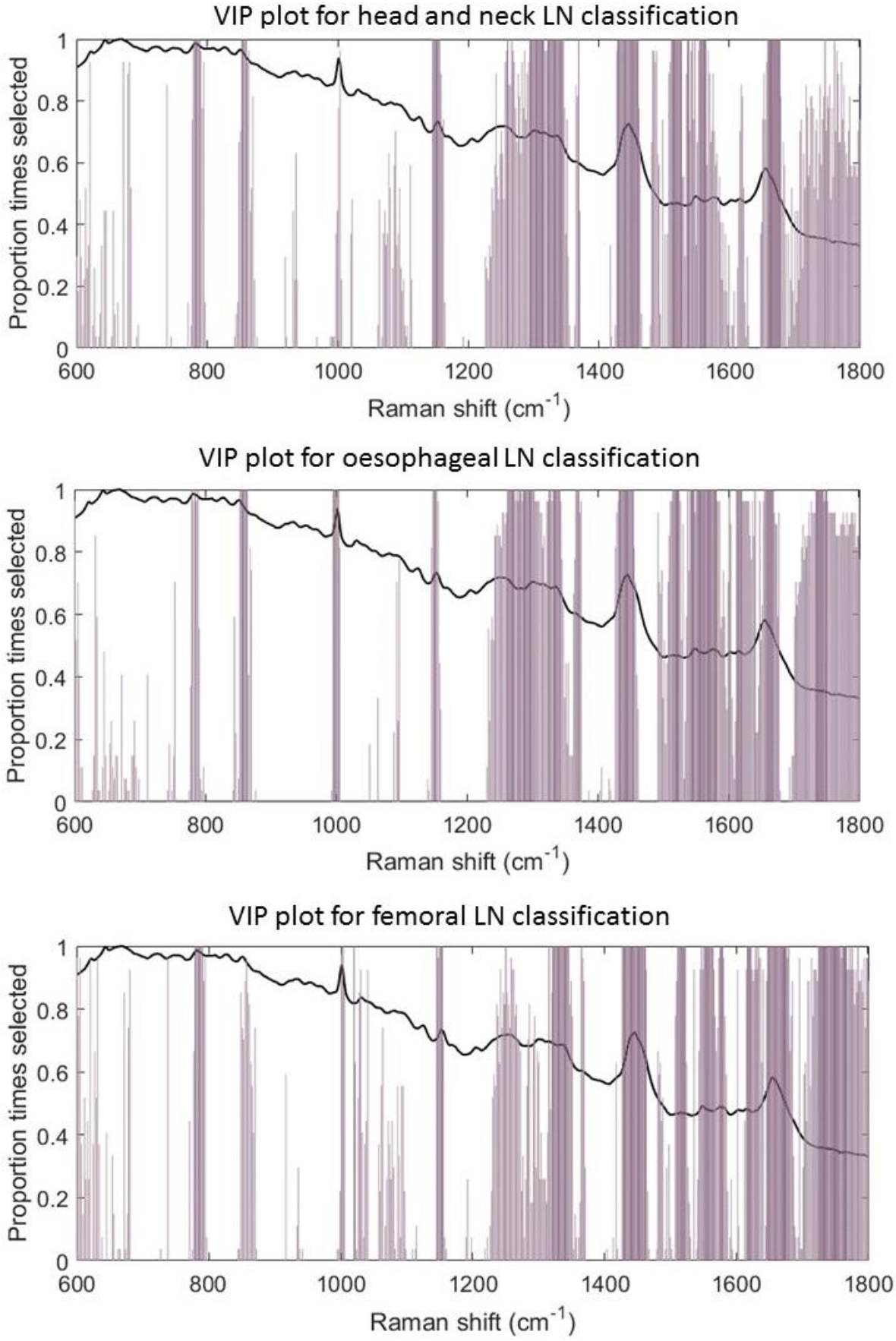

Figure 4.67- VIP plots to show the proportion of times specific Raman shifts were selected during PLSDA cross validation.

Table 4.27- Table of bands from VIP that were selected 100% of the time for the PLSDA cross validation.

Bands selected for head & neck classification 100% of the time		**Bands selected for oesophageal classification 100% of the time**		**Bands selected for femoral classification 100% of the time**	
779.1-787.8	Pyrimidine	779.1	Pyrimidine	779.1-790	Pyrimidine
852.7-859.1	Tyrosine	783.4	Pyrimidine	1001-1003	Phenylalanine
1146-1156	Protein	787.8	Pyrimidine	1020	
1260	Amide III	852.7-863.4	Tyrosine	1146	Protein
1266	Amide III	994.8-1003	Phenylalanine	1150-1154	Lipid, collagen
1294-1316	Lipid, collagen	1148-1154	Protein	1251	Amide III
1322-1345	Carbohydrate	1262-1272	Amide III	1316	Collagen
1369	Tryp	1276	Amide III	1322-1345	Collagen/carbohydrate
1429-1460	Protein, phospholipid	1282	Amide III/collagen	1365	Tryp
1511-1527	Carotenoid	1286	Cytosine	1427-1462	Protein, phospholipid
1538	Tryptophan	1288	Cytosine	1511-1523	Carotenoid
1544	Tryptophan	1302	Lipid/collagen	1547	Tryptophan
1547	Tryptophan	1326-1328	Nucleic acids	1551-1564	Tryptophan, nucleic acid
1555-1564	Nucleic acid	1332-1342	Carbohydrate	1575-1579	Nucleic acid
1659-1678	Amide I	1367-1371	Tryp	1617-1622	Amino acid (tyr, tryp)
1747	Lipid	1433-1454	Protein, phospholipid	1630	Amide I
1761	Lipid	1517-1525	Carotenoid	1633	Amide I
		1538	Tryptophan	1648-1676	Amide I
		1542-1547	Tryptophan	1679	Amide I
		1551-1579	Nucleic acid	1725-1758	Lipid
		1602	Amino acid	1761-1767	Lipid
		1613-1618	Amino acid	1772-1776	Lipid
		1641	Amide I	1785	Lipid
		1655-1668	Amide I		
		1734-1743	Lipid		
		1749	Lipid		

VIP identified a large number of peaks selected 100% of the time for classification from across the whole measured wavelength range. However, it only identifies them based on their importance for variability and cannot identify which peaks are characteristic of which groups. Thus, the bands selected 100% of the time were compared to the PCA loadings in order to assess which groups the peaks are more frequent in, hence leading to their importance for classification. In head and neck LN data the biochemical assignments of the peaks that are the most indicative of the group are tryptophan (1547 cm^{-1}), nucleic acids (1561 cm^{-1}) and amide I (1655 cm^{-1}). For oesophageal LN, the biochemical assignments of the peaks most suggestive of the dataset are phenylalanine (1001

cm^{-1}), amide III (1257 cm^{-1}), lipid/collagen (1302 cm^{-1}), carbohydrate (1338 cm^{-1}), protein/lipid (1367 cm^{-1}), proteins/phospholipid (1446 cm^{-1}), tryptophan (1551 cm^{-1}), nucleic acid (1581 cm^{-1}) and amino acid (1617 cm^{-1}). The most indicative peaks from femoral LN data are assigned as phenylalanine (1001 cm^{-1}), proteins/phospholipid (1446 cm^{-1}), tryptophan (1551 cm^{-1}), amino acid (1617 cm^{-1}), amide I (1655 cm^{-1}) and lipid (1745 cm^{-1}).

The ability to distinguish between the metastatic cancers that are found in the LN from the three anatomical regions was also of interest, to determine how identifiable their Raman spectra are in comparison to each other. Figure 4.68 shows the mean spectra of the three metastatic types found in the LN of the three regions. The metastatic cancers in the study were mouth and throat SCC found in head and neck nodes, AC from oesophageal cancers and cervical SCC in femoral nodes.

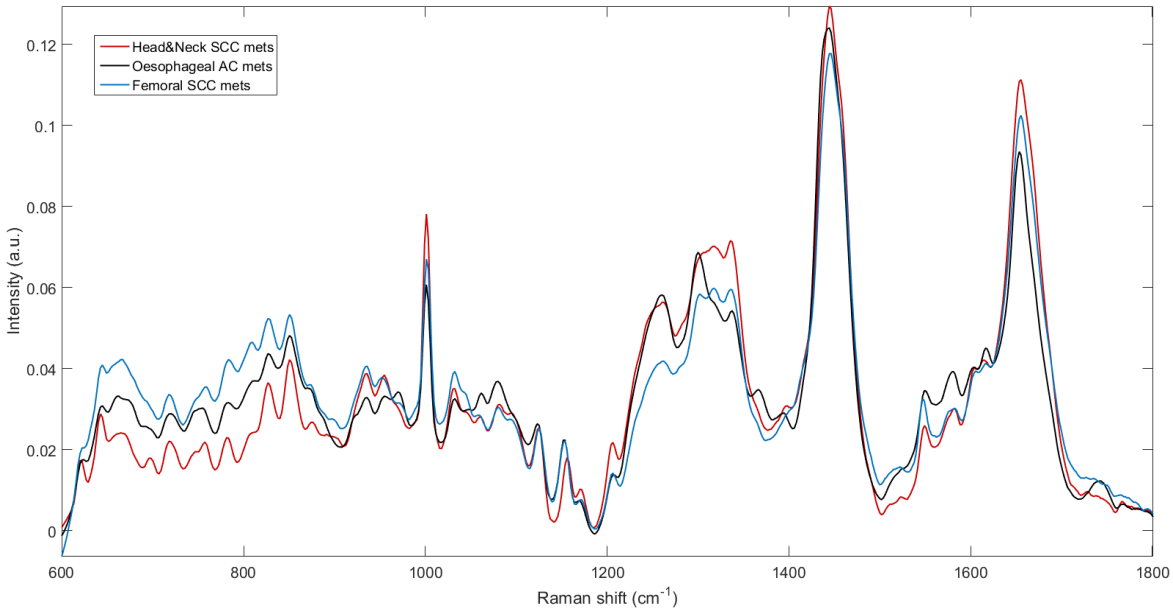

Figure 4.68- Mean spectra of SCC mets in head & neck LN (red), AC mets in oesophageal LN (black) and SCC mets in femoral LN (blue), averaged across 100, 120 and 120 measurements respectively.

PCA carried out shows unsupervised clustering of data from the three metastatic groups. The PCA scores plot in Figure 4.69 shows the results from the

top four components selected through ANOVA. PC4 and PC3 appear to separate the femoral and head and neck metastatic LN data, primarily plotted on the negative axis, from the oesophageal metastatic LN data, primarily plotted on the positive axis. Separation in PC 1 and PC5 are harder to discern, however it can be seen that metastatic data from the head and neck LN are more frequently plotted on the negative axis in PC1 and the metastatic data from the femoral LN are more heavily distributed on the positive axis in PC5.

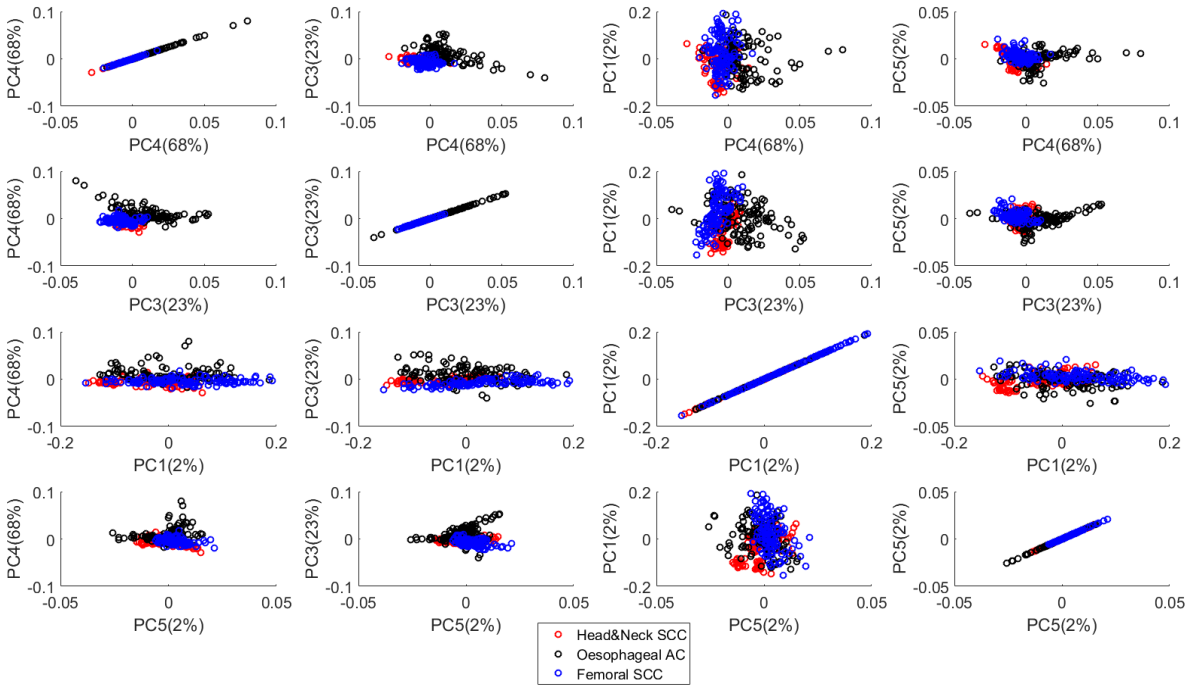

Figure 4.69- PCA scores plot showing the distribution of measurements from SCC in head & neck LN (red), AC in oesophageal LN (black) and SCC in femoral LN (blue) across the top four components.

The PCA loadings for the top four components are displayed in Figure 4.70 and the main peaks and correlating biochemical assignments are shown in Table 4.28. Through comparison of the loadings with the scores plot, the spectral peaks indicative of specific groups were identified and their biochemicals tentatively assigned. In PC4 and PC3 peaks in the negative direction, and thus suggestive of SCC metastases in femoral LN and head and neck LN, were those assigned as protein (938 cm^{-1}), lipids (1435, 1752 cm^{-1}) and amide I (1641, 1670 cm^{-1}). Those peaks in the positive direction, and thus more indicative of AC metastases in oesophageal LN, were assigned as amide III (1261, 1245 cm^{-1}), protein and

lipid (1371, 1369 cm^{-1}), carotenoid (1515, 1527, 1502 cm^{-1}), nucleic acid (1579, 1577 cm^{-1}) and amino acid (1618 cm^{-1}). The loading of PC1 seems to be affected largely by the spectral background, with a higher background in positive scores around the 600-900 cm^{-1} range and a higher background in negative scores in the 1200-1800 cm^{-1} range, suggestive of head and neck LN metastasis. In the loadings for PC5, peaks indicative of SCC metastases in femoral nodes were in the positive direction and consisted of those assigned as amino acid (675.1, 1545, 1561, 1618 cm^{-1}), DNA (750.7 cm^{-1}), lipids (1302, 1747 cm^{-1}), protein and phospholipid (1450 cm^{-1}) and nucleic acid (904, 1576 cm^{-1}).

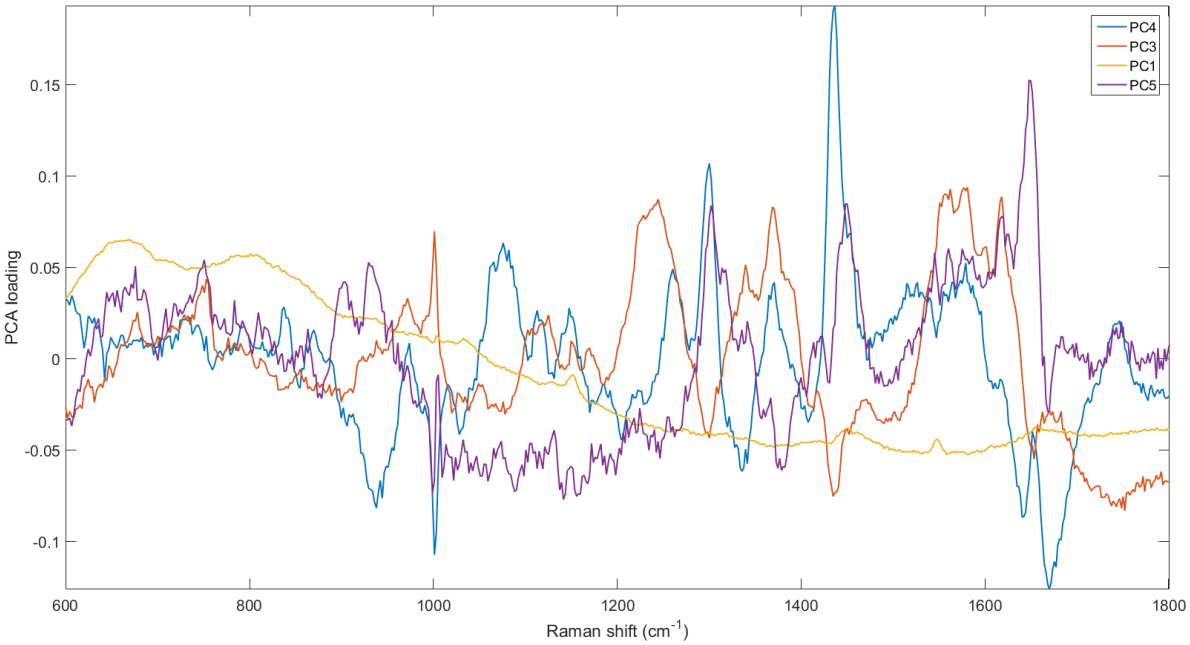

Figure 4.70- PCA loadings of the top four principal components for discriminating metastatic lymph nodes from the head and neck, oesophagus and femoral regions.

Table 4.28- Main peaks and their tentative assignments from the PC loadings.

PC4				PC3				PC1				PC5			
-ve assignments		+ve assignments		-ve assignments		+ve assignments		-ve assignments		+ve assignments		-ve assignments		+ve assignments	
938	Protein	1076	Protein, lipid	1435	Protein, phospholipid	750.7	DNA		-		-	999	Phe	675.7	Amino acid
1001	Phe	1261	Amide III	1752	Lipid	1001	Phe		-		-	1022	Phe	750.7	DNA
1336	Carb, Nucleic acid	1300	Lipid, collagen			1245	Amide III					1059	Lipid	904	Adenine
1641	Amide I	1371	Tryp			1340	Carb, Nucleic acid					1088	Lipid, DNA	929.5	Protein
1670	Amide I	1437	Protein, phospholipid			1369	Tryp					1142	Protein	1302	Lipid, collagen
		1515	Carotenoid			1502	Carotenoid					1152	Protein	1314	Collagen
		1527	Carotenoid			1577	Nucleic acid					1379	Tryp	1450	Protein, phospholipid
		1579	Nucleic acid			1618	Amino acid					1670	Amide I	1545	Tryp
		1747	lipid											1561	Tryp
														1576	Nucleic acid
														1618	Amino acid
														1648	Amide I
														1747	Lipid

PLSDA and cross validation was carried out on the data. The PLSDA scores plot, Figure 4.71, shows the clustering of the groups using the first three PLS components. It appears that the oesophageal AC and femoral SCC data are more delineated from each other than the head and neck SCC data, especially in PLS3. The sensitivities and specificities were calculated from the cross validation using a majority wins approach for the data per lymph node. Sensitivities and

specificities were calculated between 60-67% and 73-91% for metastatic lymph nodes from the three anatomical regions studied (Table 4.29).

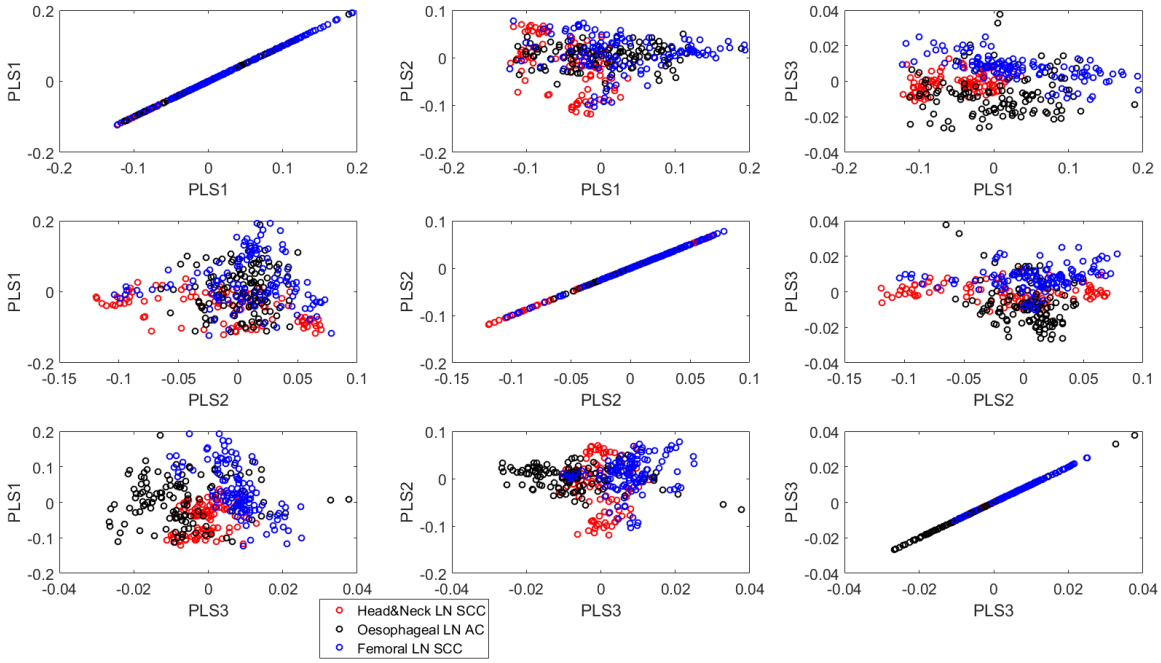

Figure 4.71- PLSDA scores plot for the first three components, showing the distribution of head & neck SCC (red), oesophageal AC (black) and femoral SCC (blue) measurements.

Table 4.29- Confusion matrix for the classification of head & neck SCC, oesophageal AC and femoral SCC LN data.

		True group		
		Head & Neck SCC	**Oesophageal AC**	**Femoral SCC**
Predicted group	**Head & Neck SCC**	3	1	1
	Oesophageal AC	0	4	1
	Femoral SCC	2	1	4
Total LN		5	6	6
Sensitivity		0.6000	0.6667	0.6667
Specificity		0.8333	0.9091	0.7273

VIP was used to identify the important spectral bands for delineating between the three metastatic groups. The VIP plot, in Figure 4.72, highlights the bands used for cross validation and the proportion of the times those bands were selected. Those that were selected 100% of the time were tentatively assigned

biochemicals and are displayed in Table 4.30. These bands were compared to the PCA weights, in order to assess which peaks are primarily indicative of the certain groups.

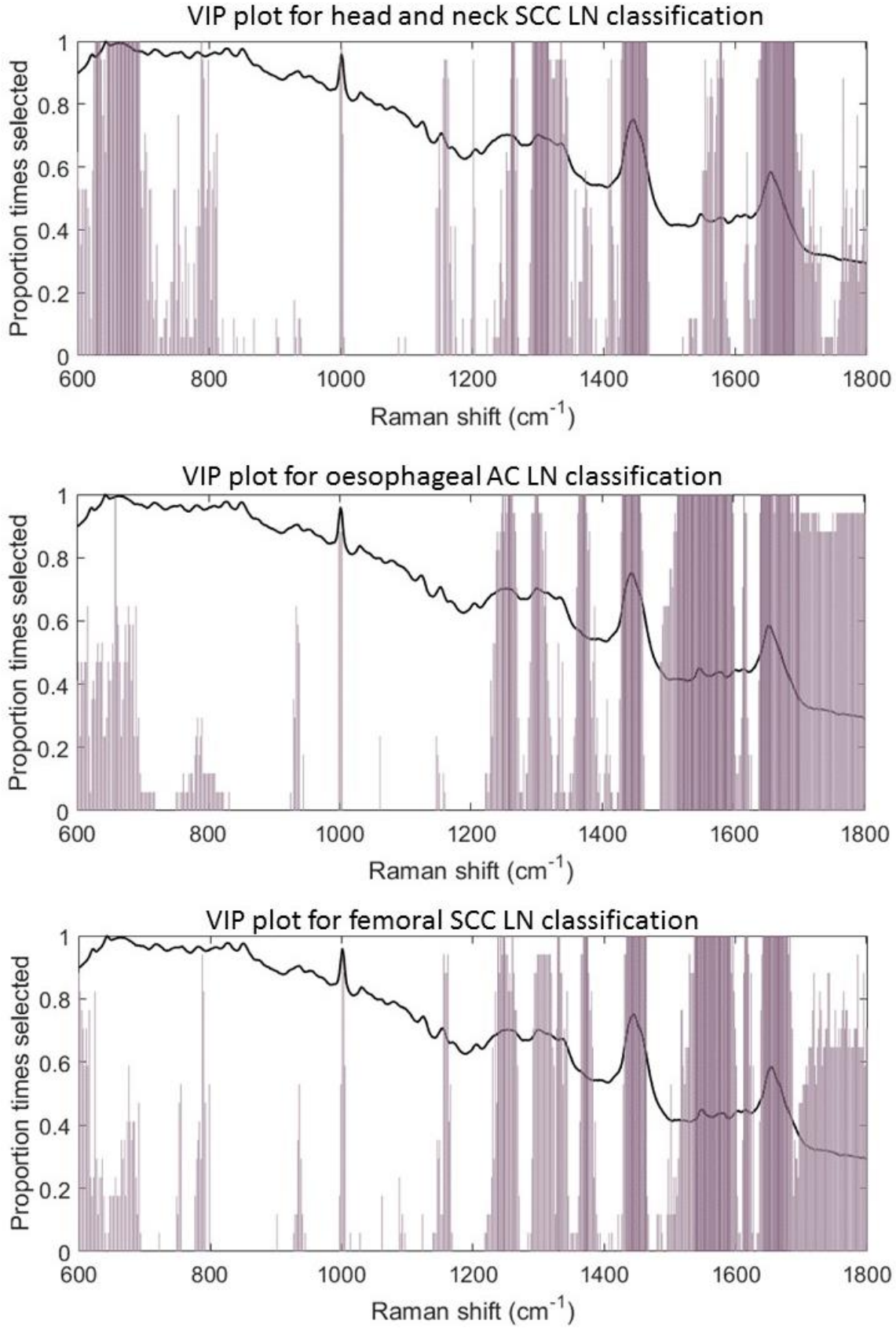

Figure 4.72- VIP plot highlighting the spectral bands that were selected for classification of the head & neck SCC, oesophageal AC and femoral SCC.

The peaks that are mainly suggestive of head and neck SCC are those assigned as phenylalanine (1001 cm^{-1}), carbohydrate and nucleic acid (1336 cm^{-1}) and amide I (1655 cm^{-1}). Those that are indicative of oesophageal AC are assigned as amide III (1261 cm^{-1}), collagen and lipid (1300 cm^{-1}), protein and lipid (1367 cm^{-1}), proteins and phospholipids (1445 cm^{-1}), carotenoid (1523 cm^{-1}) and nucleic acid (1581 cm^{-1}). The spectral peaks that are most indicative of the femoral SCC are assigned as carbohydrate and nucleic acid (1336 cm^{-1}), protein and phospholipid (1446 cm^{-1}), tryptophan (1547 cm^{-1}), nucleic acid (1581 cm^{-1}), amino acid (1617 cm^{-1}) and amide I (1655 cm^{-1}).

Table 4.30- Table of bands selected 100% of PLSDA cross validations with their tentative assignments.

Bands selected 100% of times for head & neck SCC classification		**Bands selected 100% of times for oesophageal AC classification**		**Bands selected 100% of times for femoral SCC classification**	
626.6-633.4	Phe	657.9	Amino acid	1239	Amide III
644.5-693.4	Tyr, amino acid, glycine	1249	Amide III	1243	Amide III
787.8	Pyrimidine base	1253	Amide III	1247-1249	Amide III
1001	Phe	1256-1262	Amide III	1260	Amide III
1260-1264	Amide III	1298-1302	Lipid, collagen	1330-1332	Carb, nucleic acid
1292-1316	Lipid, collagen	1365-1375	Tryp	1369-1375	Tryp
1336	Carb, nucleic acid	1431-1458	Protein, phospholipid	1435-1464	Protein, phospholipid
1429-1466	Protein, phospholipid	1515-1598	Carotenoid, tryp, nucleic acid	1538-1592	Tryp, nucleic acid
1577-1579	Nucleic acid	1617	Amino acid	1596	Amino acid
1641-1690	Amide I	1643-1661	Amide I	1615-1618	Amino acid
		1665-1687	Amide I	1643-1679	Amide I
		1690	Amide I	1683	Amide I
		1696-1700	Amide I		

The metastatic data can be grouped into both AC, found in the oesophageal LN, and SCC, found in the head and neck and femoral nodes. The ability to differentiate between these two cancer types was of interest in assessing the capability to distinguish between two types of metastatic carcinomas. PLSDA was carried out again as a two group model, grouping the metastatic femoral and head and neck LN data as one group, and the metastatic oesophageal LN as

another. Cross validation resulted in sensitivities and specificities between 83-91% using a majority wins approach per node for classification of the two metastatic carcinoma types. Table 4.31 displays the confusion matrix for the cross-validation of the PLSDA model.

Table 4.31- Confusion matrix for the classification of AC and SCC data.

		True Group	
		AC	**SCC**
Predicted Group	**AC**	5	1
	SCC	1	10
Total LN		6	11
Sensitivity		0.8333	0.9091
Specificity		0.9091	0.8333

Discussion

The initial finding of the study was that there are biochemical differences between benign lymph nodes from different regions that lead to spectral variations. The three lymph node regions included in this study were head and neck, oesophagus and femoral. The lymph nodes from these regions drain different locations of the body. The head and neck drain lymph found in the head, neck and upper oesophagus. The oesophageal lymph nodes primarily drain the mid portion of the oesophagus[260] and other upper abdominal regions. The femoral lymph nodes drain lymph from the lower portion of the abdomen and upper legs. The anatomy of lymph nodes are supposedly the same throughout the body. Therefore, the regions they drain may be responsible for the biochemical differences observed in the spectral measurements and analysis of the non-cancer LN from the different regions. This could suggest that combined LN models, like that in the previous chapter, may be less accurate than independent models for specific regions. The peaks at 1260 cm^{-1} and 1336 cm^{-1} assigned as amide III, appears to be more distinctive of oesophageal lymph node spectra, seen across the component loadings, and also highlighted as important by VIP for the identification of measurements belonging to this group. The peak at 1655 cm^{-1}, assigned as amide I, however seems to be more characteristic of femoral and head and neck lymph node measurements. Amides are compounds derived from fatty acids where the carboxyl group (COOH) is replaced by an amino group (NH_2). The differences between the amide groups will be as result of the structure of the molecule which leads to various Raman shifts and thus detection of varying concentrations between the three lymph node region groups.

The peaks that appear to be more indicative of femoral lymph nodes are 1745 cm^{-1} assigned as lipid and 1437 cm^{-1} assigned as protein and phospholipid. The majority of the lymph node samples had fat deposits seen either microscopically or through their Raman spectra. However, the results from the analysis suggest that the lipid content in the femoral lymph nodes was slightly higher in order to delineate them from the other lymph node regions. Isolated fat cells and small areas of adipose tissue are often found in lymph nodes after childhood and lymph node tissue can even be replaced by adipose tissue in the elderly, obese or as a result of chronic infection, preserving very little of the

original node structure. It is common for any lymph nodes to show a degree of lipomatosis. However, inguinal and axillary lymph node regions have the highest fat abundance.[261] The inguinal lymph node region include those that are femoral and their high fat abundance may explain the higher detected levels in their Raman measurements. Weiss *et al*.[262] explain that lipid lymphadenopathy, caused by lipid deposition from outside the lymph node, is often associated with tumour tissue and commonly occur in abdominal nodes. Furthermore, Meksiarun *et al*.[263] showed the difference in visceral and subcutaneous adipose tissue and highlight the main variation around the spectral regions 974, 1230 and 1655 cm^{-1}. Therefore, alterations amongst the lipid peaks may be influenced by this difference between fat types across different anatomical regions.

As well as the amide I peak, head and neck lymph node measurements also appear to have higher intensities of peaks assigned as amino acids. In particular tryptophan, with assigned peaks at 1559 and 1562 cm^{-1}, were found to be characteristic of the lymph nodes from this group. Amino acids play an important role in immunity aiding the activation of B cells, T cells, natural killer cells and macrophages, regulating redox state, lymphocyte proliferation and gene expression and producing antibodies and cytokines. The immune system relies heavily on the availability of amino acids for the production of proteins and polypeptides as well as individual molecules that control metabolites.[264] Amino acids can also be used for fuelling the immune system and the production of new cells either directly or through conversion to other amino acids or glucose.[265] Tyrosine is directly involved in phosphorylation and the production of neurotransmitters, such as dopamine, and melatonin, an antioxidant, and its precursor phenylalanine is responsible for the activation of Tetrahydrobiopterin. Tryptophan, another amino acid, produces inhibitors along with serotonin, a neurotransmitter, and melatonin, an antioxidant.[266] It is a possibility that head and neck nodes require higher levels of specific amino acids in their lymph nodes, maybe due to their location, which could explain why these peaks in the spectra tend to be more indicative of head and neck nodes.

Good cross-validation sensitivities and specificities of the three group PLSDA model, further supported characteristic differences between the measurements of the lymph node tissue from the three regions.

The metastatic lymph node measurements from the three regions also showed clustering in their PCA and PLSDA score plots. Through analysis of both the scores plots and loadings, it can be seen that many peaks responsible for variation between the measurements of the metastatic lymph node region are the same as those that were most responsible for variation between the regions of the non-cancer samples. For example the amide I peaks (1641, 1670 cm^{-1}) are apparently more indicative of metastases in head and neck and femoral nodes, whilst the amide III (1261, 1245 cm^{-1}) appears to be more characteristic of the metastatic oesophageal LN. Lipids (1435, 1752, 1302 and 1747 cm^{-1}) also seem to be present in higher concentrations in metastatic femoral nodes than the other regions.

Peaks assigned as carotenoids (1515, 1527 and 1502 cm^{-1}) were also associated more with metastatic oesophageal LN than those from the other regions. Carotenoids are pigmented compounds that are only synthesised by plants and micro-organisms and humans obtain them through diet. A higher level of carotenoids was also seen in the non-cancer lymph nodes where both femoral and oesophageal nodes were observed to have higher levels than head and neck lymph nodes. This implies that the differences in carotenoid content is not isolated to the development of carcinoma tissue in the nodes. An explanation for this higher carotenoid level in oesophageal lymph nodes could be as a result of a positive relationship with the lipid content. Carotenoids are known to be fat soluble and rely on dietary fat for absorption.[267] Therefore, as the oesophageal lymph node samples comprise a larger amount of adipose tissue, the number of carotenoids may also be higher as a result. Another reason for the observed regional differences may be the proximity of femoral and oesophageal lymph nodes to the intestines, when compared to the head and neck nodes, where carotenoids are absorbed from food sources and initially enter the bloodstream.

A peak assigned as amino acid at 1618 cm^{-1} was identified as being indicative of head and neck nodes in non-cancer samples study, yet indicative of oesophageal samples in the metastatic LN study. This shows that not all differences between the metastatic samples must be similar to those between the

non-cancer samples and suggests that variations may be altered in metastatic tissue as a result of tumour tissue and not just lymphoid tissue.

Cross validation shows that the metastatic node measurements are analytically more similar to each other than those from non-cancer nodes seen through their cross-validation results. This may suggest that metastasis in lymph nodes, regardless of location, has dominant and related biochemicals that cause their Raman signature to resemble each other. As the biochemicals that lead to the differences between the LN regions are very similar in both the non-cancer and cancer samples, yet with a drop in sensitivity to predict, there is an implication that metastasis leads to a reduction in variation between LN from different regions. However, all the metastases in this study were carcinomas, originating from epithelial tissue. Therefore, there may be a higher level of homogeneity between epithelial tissues than lymphoid. As a result, the reduction in variation would probably arise due to the introduction of new biochemicals from the epithelial cells, masking the important identifying features between lymph node regions. It would be of interest to measure a variety of healthy epithelial tissue in order to spectrally assess any similarity to each other and whether this could be responsible for lack of region identification, or whether metastatic tissue as a whole is more similar to each other and causing this reduction in variation of lymph node regions.

Both the femoral and head and neck lymph node metastases were squamous cell carcinomas. Similarities between these groups may have resulted in decreased sensitivities for the identification of the different lymph node regions when invaded with metastasis. When grouping the squamous cell carcinoma data, from both the head and neck and femoral node measurements, together as one pathology and leaving the adenocarcinoma as another, the classification sensitivities and specificities greatly improve. It is understood that the reduction of the model to just two groups will contribute to this improvement as a result of the 50% baseline classification of 2-group models. However, the greater sensitivities may also be due to similar biochemicals being present between the two SCC groups that allow its better identification from AC measurements when grouped together. Such as, the SCC groups both have a more intense 1336 cm^{-1} peaks assigned as a combination of nucleic acid and carbohydrate. It may also

be possible though, that non-cancer femoral and head and neck lymph nodes already share more similarities, such as a stronger amide I peak, than with oesophageal lymph nodes and that is what is leading to the better classification. A further investigation using samples with both AC and SCC metastases from the three lymph node regions would be a more accurate way to determine whether SCC and AC metastases can truly be identified from one another. It is especially of interest in determining the ability to clinically diagnose a patient more accurately, or even potentially primary site identification, rather than grouping metastasis or cancer as a whole.

Conclusion

It has been determined that spectral changes exist between the regions of head and neck, oesophageal and femoral lymph nodes in both non-cancer and metastatic samples. The classification ability for the various lymph node regions is higher in the non-cancer sample measurements than those of the metastatic. However, the biochemicals that appear to cause the differences between the three lymph node regions, regardless of pathology, remain similar. These variations between the regions appear to be primarily because of amide I and III, lipids, amino acids and carotenoids peak intensities. The ability to potentially identify SCC from AC metastasis was also observed and would be extremely useful for clinical diagnosis. Further studies with larger sample sizes and a larger range of pathologies would be required in order to assess the source of variability between the metastatic groups however.

These findings suggest that lymph nodes from different locations within the body may contain too many varying biochemical features to allow them to be treated as one single tissue group. When analysing lymph nodes it might be more appropriate to consider individual regions, comparing like samples, for discrimination and identification of pathological groups.

Chapter 5: Raman Analysis of Primary and Secondary Cancers in Head and Neck Lymph Node Tissue

5.1 Results

Inflammation of head and neck lymph nodes are more noticeable due to their superficial anatomy in many cases. Whilst inflammation is often the result of benign conditions, it can also be caused by primary and secondary malignancy, and in such cases the node may also feel harder than usual. Lymphomas and metastases are amongst the many diseases frequently diagnosed in the head and neck. From the samples collected for this study, the head and neck lymph nodes contained the widest array of pathologies. This part of the research spectrally examined 15 µm sections of head and neck lymph nodes to assess detectable biochemical differences between diseased lymph nodes and those without any diagnosed pathological abnormalities. An area of the section was mapped after reviewing the adjacent H&E with the pathologist to identify key areas, the Raman images were then examined and the spectra from a region of interest (ROI) was used for analysis.

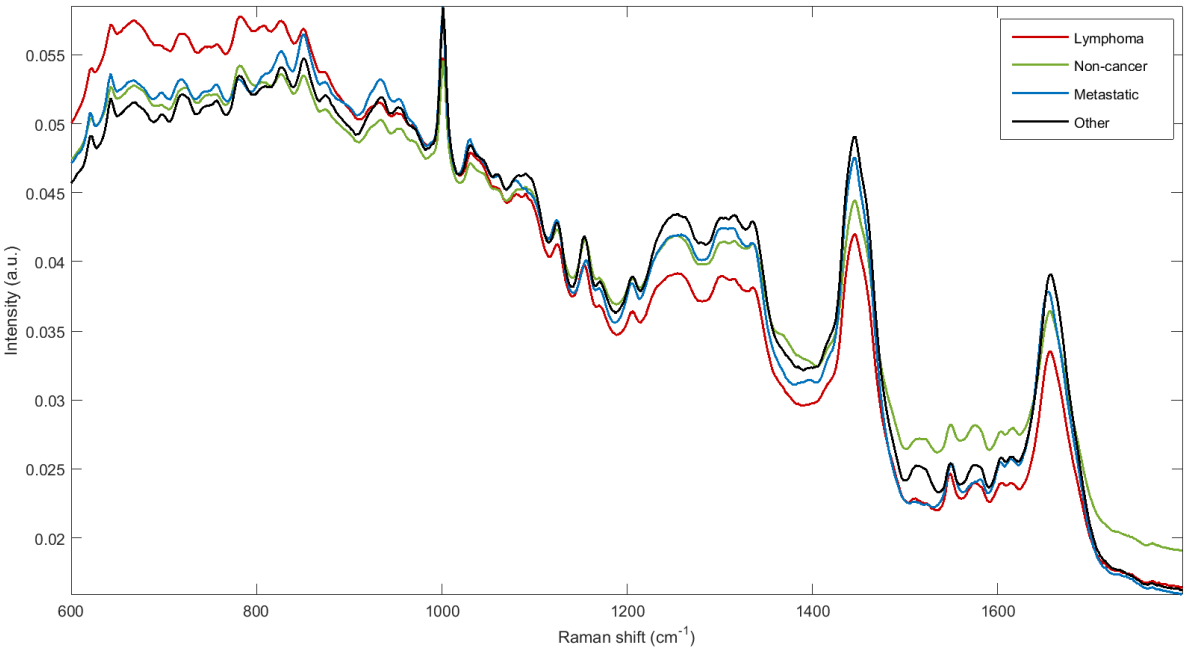

Figure 5.73- Mean spectra of lymphoma, metastatic, other and non-cancer maps, averaged across 33201, 12120, 5151 and 53879 acquisitions respectively.

The average spectra of the four histological groups, *lymphoma*, *metastatic*, *other* and *non-cancer*, are displayed in Figure 5.73. The most visually evident difference between the spectra appears to be in the 1515 cm^{-1} region assigned as carotenoid, where the peak is much less intense in the averaged *metastatic* and *lymphoma* samples' measurements than in the *non-cancer* and *other* spectra.

The intensity of these spectral peaks seem to contribute the most to differences seen visually, rather than spectral shift which may be subtly present and require further multivariate analysis to detect. As expected, the two benign groups, *non-cancer* and *other*, are apparently more spectrally similar to one another. The *metastatic* mean spectrum looks the most dissimilar to the other groups, whilst the averaged *lymphoma* spectrum shares similarities with and exhibits differences between all the groups over the measured wavenumbers. Multivariate analysis was then employed to mathematically analyse further variations between the dataset.

To provisionally assess if cancer could be distinguished from non-cancer in this investigation, the *metastatic* and *lymphoma* data were classed as one group, *cance*r. The *non-cancer* and *other* data was classed as another group, *non-cancer*. A two group model is faster to initially identify differences between the groups and assess classification capability.

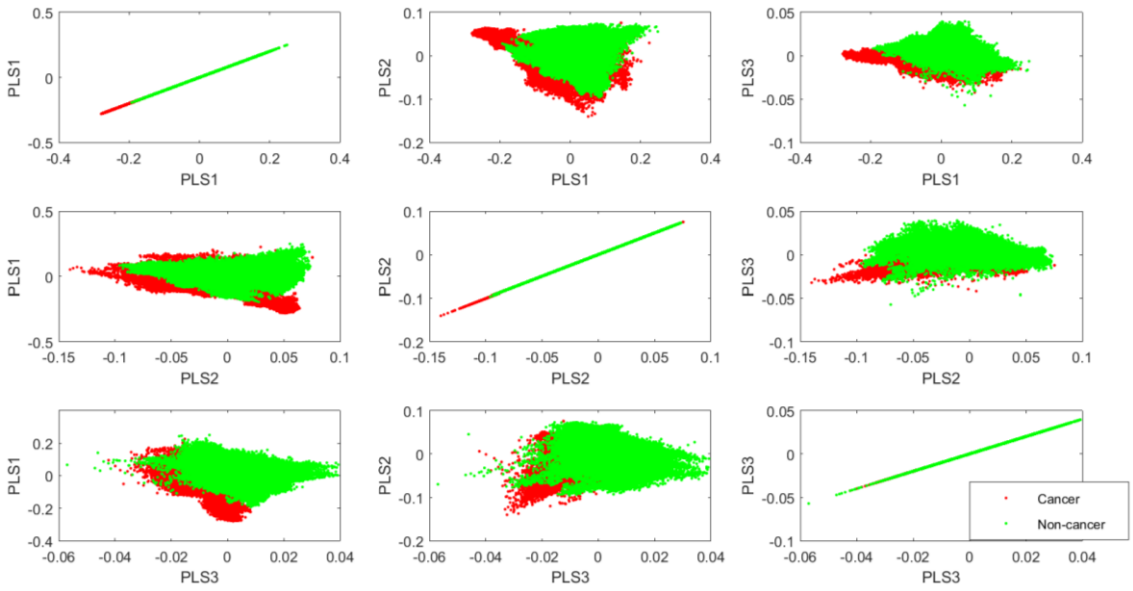

Figure 5.74- PLSDA scores plot for cancer (45321 acquisitions) and non-cancer (59030 acquisitions) measurements

The PLDA scores plot, Figure 5.74, shows the clustering of the data according to the top three components, of the malignant groups and benign groups combined. The *cancer* group consists of *lymphoma* and *metastatic*, and the *non-cancer* is both the *other* and *non-cancer* groups. Whilst there is no distinct separation between the data, the groups cluster with their own more readily. In the first two components the cancer scores are more negative whilst the non-cancer scores more positive. The third component also clusters the data the same with a slight few outliers in the non-cancer group. The lack of distinction between the groups could be due to the combination of a wide variety of pathologies within both the groups. This can cause difficulty for differences between the groups to be identifiable and representative of the whole group. A leave-one-sample-out cross-validation was carried out on the PLSDA model and a sensitivity of 81% and a specificity of 90% was calculated for identifying *cancer* tissue from *non-cancer* tissue in the lymph nodes using a majority wins approach (Table 5.32).

Table 5.32- Confusion matrix from the cross-validation of the PLSDA model for cancer and non-cancer LN measurements using a leave-one-sample out majority wins approach.

		True group	
		Cancer	**Other & Non-Cancer**
Predicted group	**Cancer**	18	4
	Non-Cancer	4	36
	Total Samples	22	40
	Sensitivity	0.8182	0.9000
	Specificity	0.9000	0.8182

The group labelled *other*, which are samples that were found to be reactive as a result of a disease other than cancer were then removed from the dataset and the PLSDA model was run again. Cross validating this model produced a sensitivity and specificity of 77% and 89% respectively (Table 5.33). Surprisingly, one more *cancer* lymph node was incorrectly identified as *non-cancer* in this PLSDA model than in the previous. Also, in the previous model none of the three *other* lymph nodes were identified as *cancer* but it was instead four of the diagnosed *non-cancer* lymph nodes.

Table 5.33- Confusion matrix from the cross-validation of the PLSDA model for cancer and non-cancer LN measurements using a leave-one-sample out majority wins approach excluding the 'other' group.

		True group	
		Cancer	**Non-Cancer**
Predicted group	**Cancer**	17	4
	Non-Cancer	5	33
	Total Samples	22	37
	Sensitivity	0.7727	0.8919
	Specificity	0.8919	0.7727

Though the model without the *other* group did not have quite as high sensitivities and specificities than the one that included it, the *non-cancer* group should theoretically provide a more representative sample set of non-diseased

tissue as no abnormal pathology was identified, and therefore this model was further analysed.

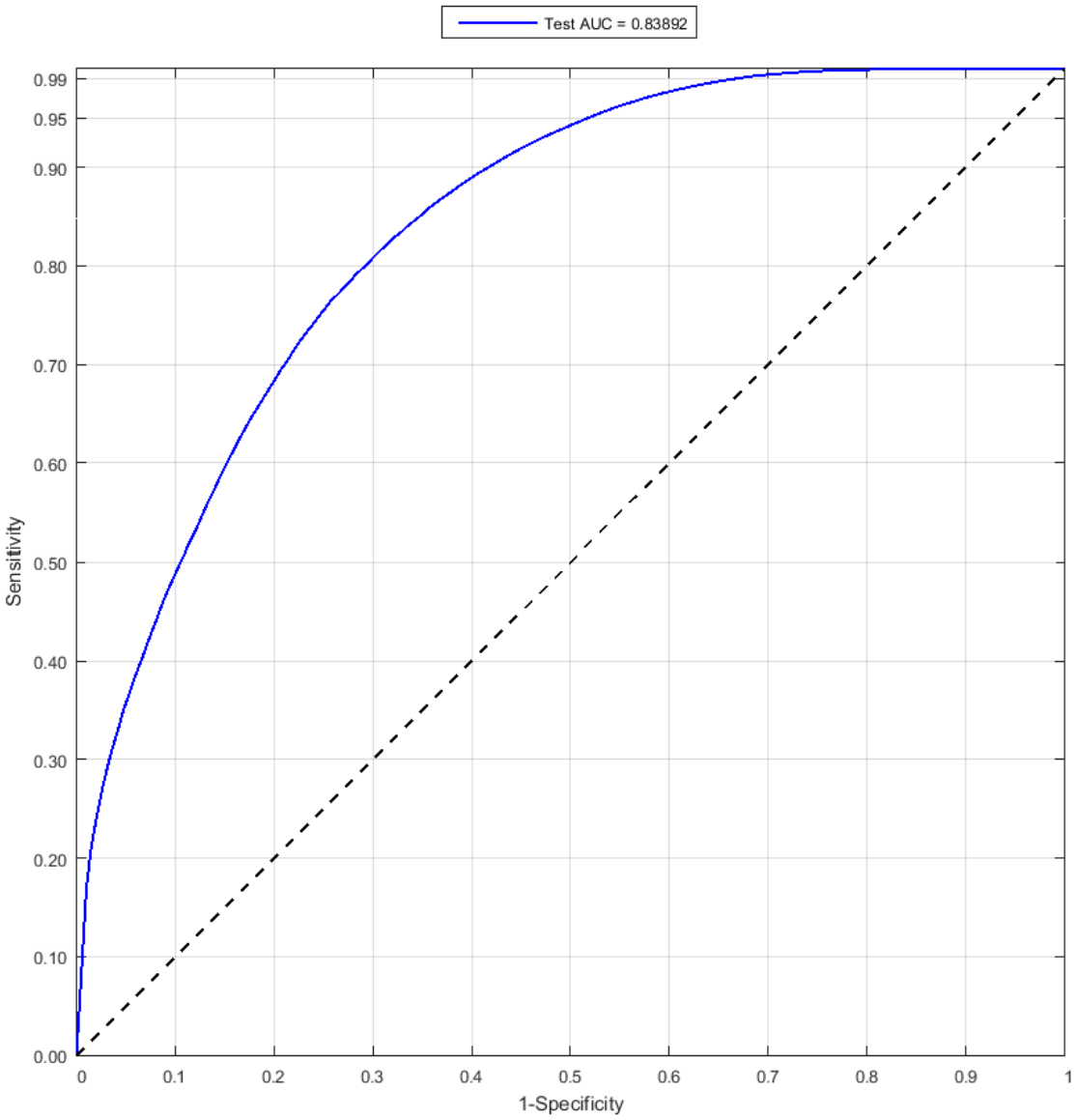

Figure 5.75- ROC curve for the PLSDA model delineating cancer from non-cancer.

The ROC curve in Figure 5.75 shows the relationship between the sensitivity and specificity of the PLSDA model for identifying *cancer* tissue from *non-cancer* tissue by sample. It shows how the specificity or sensitivity can be compromised in order to improve the other. For example, it may be more important to have a higher sensitivity at the expense of a lower specificity when identifying *cancer* tissue, so not to incorrectly diagnose a patient as healthy. In this model, to achieve a sensitivity of 90%, 95% and 98% a specificity of 58%, 48% and 38% respectively would follow. In order to get a 100% sensitivity where

you would expect no false negatives, a specificity of just 28% would occur with this model.

The PLSDA weights spectra are displayed in Figure 5.76 for the first three PLS components. The weight from the first component is clearly affected by background and therefore negative and positive peaks were categorised based on their direction as oppose to their peak weight.

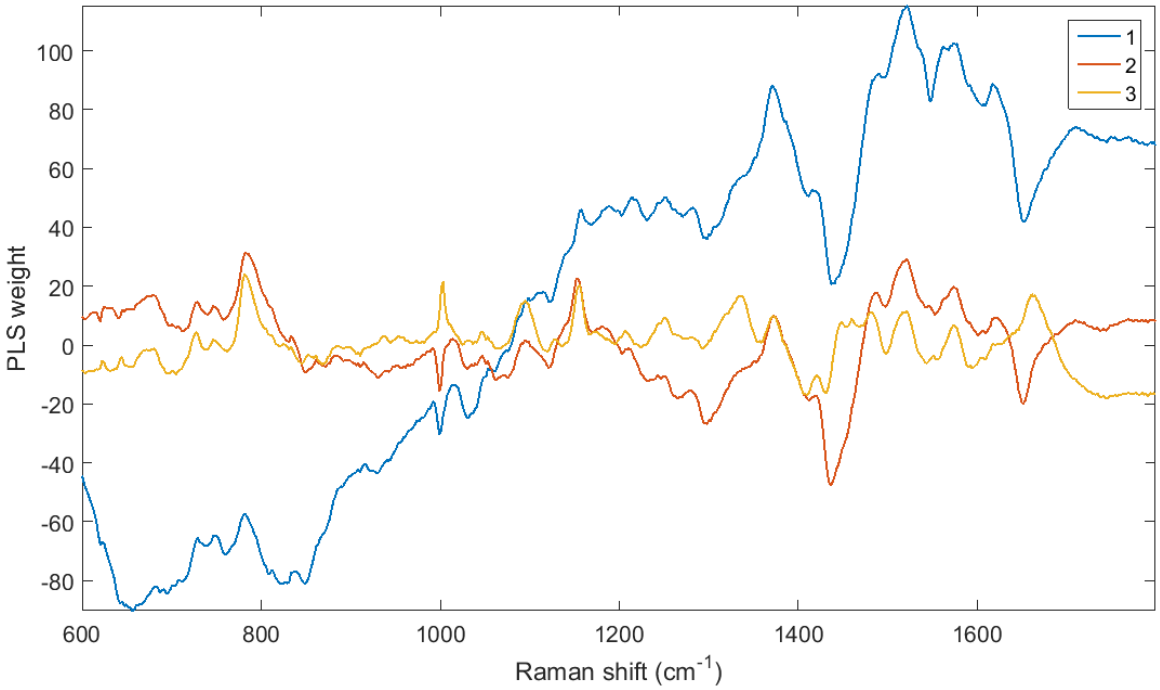

Figure 5.76- PLSDA component 1, 2 and 3 weight spectra.

Across the three components, peaks in the negative direction, and thus more indicative of the *cancer* lymph node measurements, include: peaks assigned to DNA and nucleic acid (1498 cm^{-1} and 1594 cm^{-1}) amino acids (620, 640, 695, 704, 848, 999, 1030, 1545, 1553, 1609 cm^{-1}) lipids and phospholipids (1061, 1076, 1121, 1437 cm^{-1}, amide I, 1652 cm^{-1}) and proteins including collagen (1076, 1121, 1298, 1412, 1437 cm^{-1}). Peaks in the positive direction, and thus more indicative of the *non-cancer* lymph node measurements, include: those assigned to DNA and nucleic acids (727, 746, 781, 811, 1094, 1487, 1574 cm^{-1}) amino acids (669, 682, 1188, 1206 1214, 1373, 1562, 1622 cm^{-1}) amide III (1253, 1338 cm^{-1}) glycogen (1047 cm^{-1}) and carotenoid (1521 cm^{-1}). A full list of the main peaks and their tentative biochemical assignments are displayed in Table 5.34.

Table 5.34- List of positive and negative peaks from the PLSDA component weights.

PLS component 1				PLS component 2				PLS component 3			
-ve		+ve		-ve		+ve		-ve		+ve	
693.5	Cystine	726.6	DNA	620	Phe	624.4	Phe	695.7	Cystine	622.2	Phe
824.7	Nucleic acid	746.3	DNA	640.1	Tyr	680.2	Aino acid	704.5	Cystine	644.6	Tyr
848.4	Tyr	781.3	Pyrimidine base	999	Phe	726.6	DNA	1294	Collagen	669.1	Amino acid
999	Phe	811.7	Phosphodiester	1061	Lipid	746.3	DNA	1410	Protein	682.4	Amino acid
1030	Phe	833.3	Protein	1076	Lipid, protein	781.3	Pyrimidine base	1431	Phospholipid, protein	726.6	DNA
1061	Lipid	1094	DNA	1121	Lipid, protein	833.3	Protein	1498	DNA	746.3	DNA
1076	Protein, lipid	1156	Lipid	1298	Collagen	1047	Glycogen	1544	Tryp	781.3	Pyrimidine base
1121	Lipid, protein	1188	Protein, amino acid	1412	Protein	1094	DNA	1557	Tryp	852.7	Tryp
1298	Collagen	1214	Protein, amino acid	1437	Phospholipid, protein	1156	Lipid	1594	Nucleic acid	938	Protein
1437	Protein, phospholipid	1253	Amide III	1498	DNA	1188	Protein, amino acid	1609	Amino acid	1003	Phe
1496	DNA	1371	Tryp	1544	Tryp	1214	Protein, amino acid			1047	Glycogen
1547	Tryp	1521	Carotenoid	1553	Tryp	1253	Amide III			1094	DNA
1609	Amino acid	1562	Amino acid	1602	Amino acid	1373	Tryp			1127	Lipid, protein
1652	Amide I	1574	Nucleic acid	1652	Amide I	1487	DNA			1156	Lipid
						1521	Carotenoid			1206	Protein, amino acid
						1574	Nucleic acid			1253	Amide III
						1622	Amino acid			1338	Amide III
										1373	Tryp
										1448	Phospholipid, protein
										1460	Phospholipid, protein

										1481	DNA
										1521	Carotenoid
										1574	Nucleic acid
										1661	Amide I

VIP was carried out to identify the wavenumber regions in the spectrum that contribute to the most variation between the groups. The VIP plot, Figure 5.77, shows the proportion of times certain wavenumbers were chosen based on their ability to aid classification of the two groups. Important peaks which are apparently more indicative of the *cancer* group measurements include: those assigned to cystine, phenylalanine, collagen, other proteins, phospholipid, lipid and amide I. Important peaks which are apparently more indicative of the non-cancer group measurements include: DNA, amide III, nucleic acid and carotenoid.

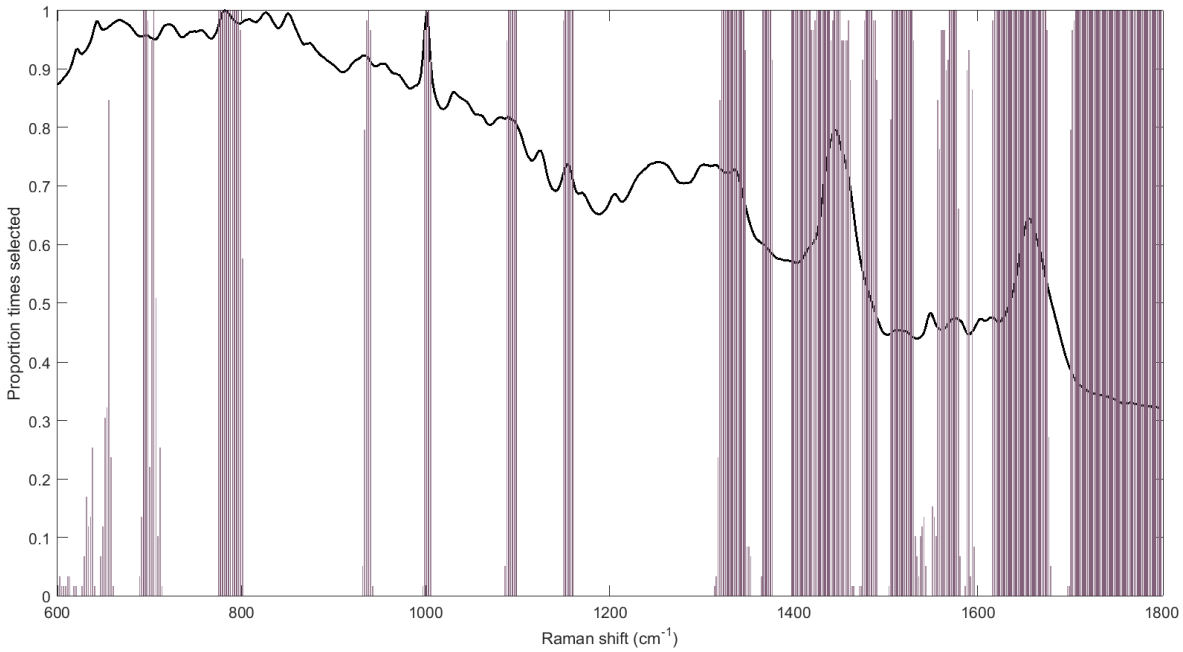

Figure 5.77 Cancer v non-cancer VIP plot with mean spectrum

Table 5.35- Table of bands selected 100% of the time for classification with their tentative biochemical assignment.

Bands selected 100% of times	**Tentative biochemical assignment**
693.5-695.7	Cystine
704.5	Cystine
774.8-796.5	Pyrimidine base
938	Protein backbone
1001-1003	Phenylalanine
1090-1099	DNA
1152-1160	Proteins
1322-1346	Collagen, amide III
1367-1375	Tryptophan, protein
1398-1418	Protein
1425-1439	Protein, phospholipid
1445-1450	Protein, phospholipid
1479-1485	DNA
1508-1528	Carotenoid
1570-1577	Nucleic acid
1618-1674	Amino acid, amide I
1707-1799	Lipid

Combining *metastatic* with *lymphoma* and *non-cancer* with the other measurements and classifying them as one pathology is flawed however. The *other* samples were still diagnosed as diseased and therefore still classified as pathological where the *non-cancer* group samples were not. Also the *metastatic* samples contained malignant cells from different organs whereas the lymphoma samples only contain malignant lymphocytes. Therefore it was essential to further analyse the *lymphoma* and *metastatic* measurements as individual classifications to ensure the differences between the two are identified and assessed. When included, the other group was also considered a separate group, however there are only three samples within the group and the pathologies are different from one another. Therefore, the results from such analyses were for interest and treated more cautiously.

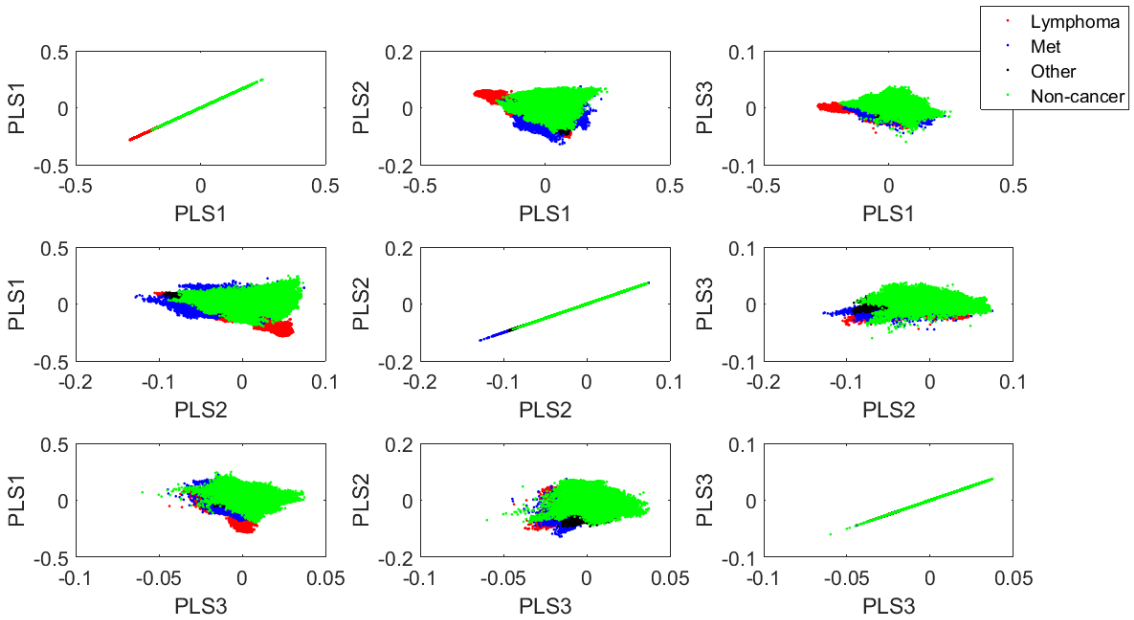

Figure 5.78- PLSDA scores plot showing the clustering of lymphoma (33201 acquisitions), metastatic (12120 acquisitions) other (5151 acquisitions) and non-cancer (53879 acquisitions) lymph nodes.

The PLSDA scores plot, Figure 5.78, shows the distribution of the *lymphoma*, *metastatic*, *other* and *non-cancer* measurements. As this is a four group model and there are also a large number of measurements involved it is difficult to discern whether certain group scores are more positive or negative. Visually it would appear that PLS component 1 separates *lymphoma* measurements, which look to be more negative, from the rest of the groups, whilst PLS component 2 separates *metastatic* and *other* measurements, which look to be more negative, from the rest of the groups. PLS component 3 looks more difficult to deduce, which is expected as the variation between groups decrease with ascending component. The weights will not be analysed in this four group model, but will be in subsequent two group models as they provide more clarity. Cross validation of the model resulted in sensitivities of 50%, 25%, 33% and 92% and specificities of 90%, 94%, 94% and 68%, for classification of *lymphoma*, *metastatic*, *other* and *non-cancer* lymph nodes from their spectra using a majority wins approach. Through analysis of the confusion matrix, Table 5.36, *lymphoma* samples are most frequently misclassified as *non-cancer*, metastatic data are more often misclassified as *lymphoma*, the samples in the *other* group are

identified as everything but *lymphoma* and the *non-cancer* group are more likely to be falsely identified as *lymphoma* or *other*.

Table 5.36- Confusion matrix from cross validation of the PLSDA model showing classification of lymphoma, metastatic, other and non-cancer data.

		True group			
		Lymphoma	**Metastatic**	**Other**	**Non-cancer**
Predicted group	**Lymphoma**	7	3	0	2
	Metastatic	1	2	1	0
	Other	1	1	1	1
	Non-cancer	5	2	1	34
	Total Samples	14	8	3	37
	Sensitivity	0.5000	0.2500	0.3333	0.9189
	Specificity	0.8958	0.9444	0.9492	0.6800

When the *other* group was combined with the rest of the *non-cancer* group, cross-validation produced slightly better sensitivities for *lymphoma* and *metastatic* lymph node classification of 57% and 38%. Sensitivities should already start higher in a three group model than a four-group model, and looking at the confusion matrix, Table 5.37, one more lymph node in both the *lymphoma* and *metastatic* group were identified correctly. The misclassifications have remained fairly similar to the four-group model.

Table 5.37- Confusion matrix from the cross-validation of the PLSDA model or classifying lymphoma, metastatic and other combined with non-cancer data.

		Actual group		
		Lymphoma	**Metastatic**	**Other & Non-cancer**
Predicted group	**Lymphoma**	8	3	3
	Metastatic	1	3	1
	Other & Non-cancer	5	2	36
	Total LN Samples	14	8	40
	Sensitivity	0.5714	0.3750	0.9000
	Specificity	0.8750	0.9630	0.6818

When removing the *other* group entirely from the analysis, the results of the cross-validation are similar to the previous model. It can be assumed from this confusion matrix, Table 5.38, that the three samples from the *other* group were classified correctly as *non-cancer* in the previous model as the only difference in this model is three less true positive identifications. This may imply that the lymph nodes classified as *other* are more similar to the lymph nodes grouped as *non-cancer*.

Table 5.38- Confusion matrix from the cross-validation of the PLSDA model for the classification of lymphoma, metastatic and non-cancer data.

		Actual group		
		Lymphoma	**Metastatic**	**Non-cancer**
Predicted group	**Lymphoma**	8	3	3
	Metastatic	1	3	1
	Non-cancer	5	2	33
	Total LN Samples	14	8	37
	Sensitivity	0.5714	0.3750	0.8919
	Specificity	0.8666	0.9607	0.6818

To better evaluate the differences between the specific pathology groups a two-group model was produced including just the *lymphoma* and *metastatic*

measurements. The PLSDA scores plot, Figure 5.79, shows the clustering of the data of the two groups. The two groups however do not show any distinct separation. The *lymphoma* scores appear slightly more negative in all the PLS components and the *metastatic* scores are more often positive for the components, apart from PLS component 2 where it appears fairly distributed across the axis and a more neutral scores on average.

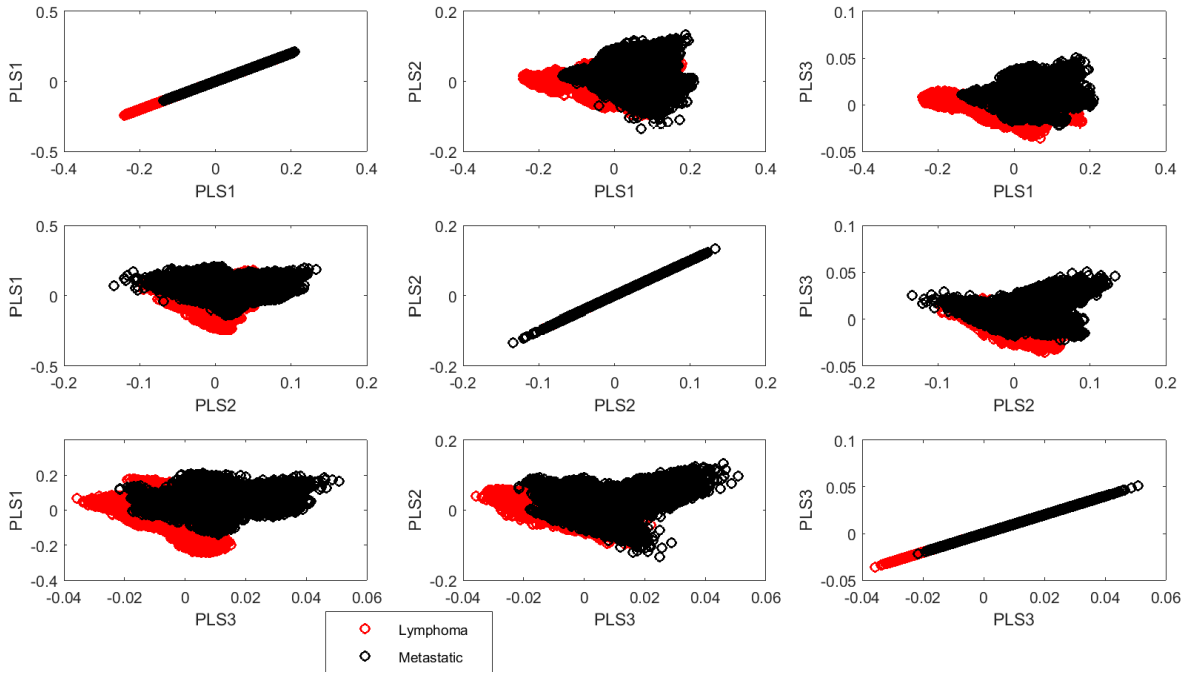

Figure 5.79- PLSDA scores plot of the lymphoma (33201 acquisitions) and metastatic (12120 acquisitions) lymph node Raman map data.

Cross-validation was performed on the PLSDA model and a sensitivity and specificity of 86% and 63% respectively for the identification of *lymphoma* measurements and 63% and 85% for *metastatic* measurements was calculated. Using these measurements and this model there is a higher ability of recognising a *lymphoma* lymph node than a *metastatic* one, the confusion matrix is displayed in Table 5.39. A slight unbalance of sample and spectrum numbers may contribute to this shift towards better *lymphoma* classification. It also supports the findings of the three and four group model where *lymphomas* were more accurately classified than *metastatic* based on their measurements.

Table 5.39- Confusion matrix for the cross-validation of lymphoma and metastatic measurements.

		Actual group	
		Lymphoma	**Metastatic**
Predicted group	**Lymphoma**	12	3
	Metastatic	2	5
	Total Samples	14	8
	Sensitivity	0.8571	0.6250
	Specificity	0.6250	0.8571

Although there is no definite separation between the two groups, the weights of the components were analysed (Figure 5.80) to evaluate if certain biochemical contributions were more indicative of specific groups. A list of the main peaks and their tentative biochemical assignments are displayed in Table 5.40. Peaks assigned as DNA and nucleic acids are overall more prevalent in the negative direction, and thus appear to be more indicative of *lymphoma* measurements, include 727, 749, 786, 833, 1092, 1487 and 1574 cm^{-1}. Also, a peak at 1257 cm^{-1} assigned as amide III and peaks at 1150, 1160 and 1513 cm^{-1} assigned as carotenoid also is apparently more indicative of *lymphoma* measurements. Peaks in the positive direction, and thus more characteristic of *metastatic* measurements, include amino acid assigned peaks at 620, 642, 696, 755, 851, 1001, 1026, 1551, 1602 and 1613 cm^{-1} and also peaks assigned as lipid and proteins at 936, 1061, 1076, 1123, 1394 and 1445 cm^{-1}.

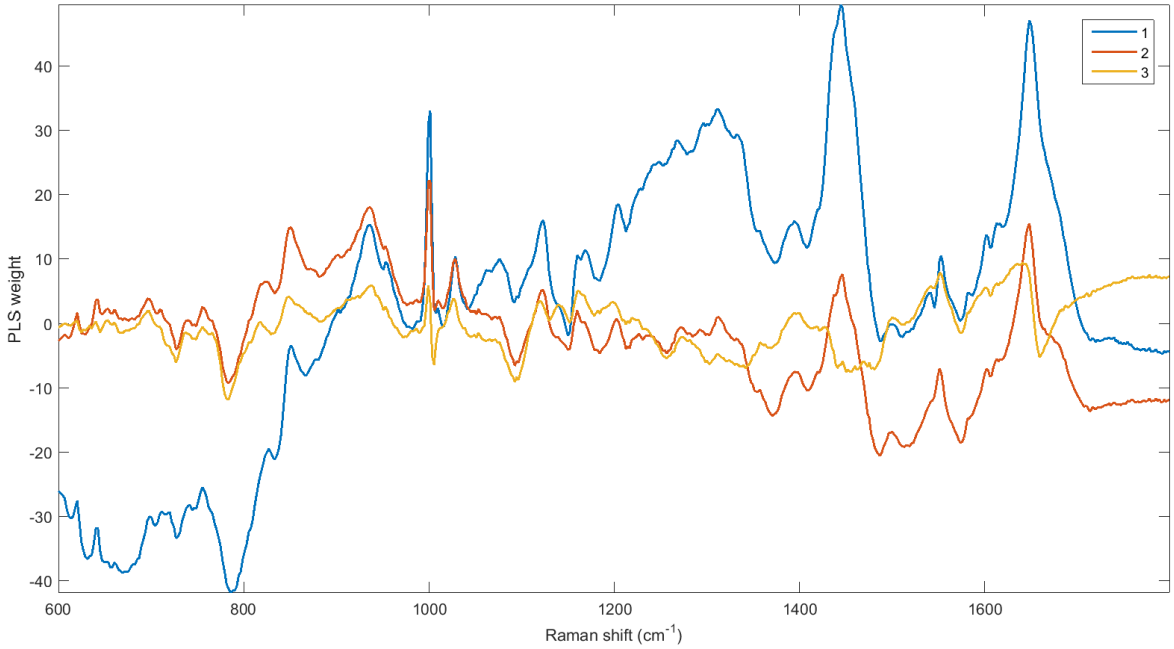

Figure 5.80- Weight spectra for the first three PLS components.

Table 5.40- List of the main negative and positive peaks for the weights of the three components with their tentative biochemical assignment.

PLS component 1				PLS component 2				PLS component 3			
-ve		+ve		-ve		+ve		-ve		+ve	
727	DNA	620	Phe	645	Tyr	620	Phe	645	Tyr	696	Cystine
749	DNA	642	Tyr	727	DNA	642	Tyr	727	DNA	848	Tyr
786	Pyrimidine	755	Tryp	784	Pyrimidine	696	Cystine	786	Pyrimidine	938	Protein
833	Nucleic acid	851	Tyr	829	Tyr	851	Collagen	829	Nucleic acid	1001	Phe
1092	DNA	936	Protein	1092	DNA	936	Protein	1005	Phe	1026	Phe
1150	Protein, carotenoid	957	HA, cholesterol	1150	Protein, carotenoid	1001	Phe	1092	DNA	1121	Lipid, protein
1487	DNA	1001	Phe	1257	Amide III	1026	Phe	1152	Protein, carotenoid	1394	Protein
1513	Carotenoid	1026	Phe	1371	Tryp	1123	Lipid, protein	1257	Amide III	1551	Tryp
		1061	Lipid	1487	DNA	1160	Protein, carotenoid	1302	Lipid, collagen		
		1076	Protein, lipid	1513	Carotenoid	1202	Protein	1481	DNA		
		1123	Lipid, protein	1574	Nucleic acid	1394	Protein	1513	Carotenoid		
		1160	Protein, carotenoid			1446	Phospholipid, protein	1574	Nucleic acid		
		1204	Protein			1551	Tryp	1659	Amide I		
		1268	Amide III			1602	Amino acid				
		1312	Collagen			1648	Amide I				
		1334	Amide III								
		1394	Protein								
		1445	Phospholipid, protein								
		1553	Tryp								
		1581	Nucleic acid								
		1602	Amino acid								
		1613	Amino acid								
		1648	Amide I								

VIP was carried out to ascertain the peaks of most importance for delineating between the two groups during cross-validation (Figure 5.81). Peaks selected 100% of the times for cross-validation are listed in Table 5.41 and include regions of DNA, amides, proteins, lipids and amino acids. Through comparison to the weights it can be deduced that DNA (724-729 cm^{-1}, 1086-1101 cm^{-1}, 1487 cm^{-1}), pyrimidine (773-801 cm^{-1}) and amide III (1257 cm^{-1}) are most important for being characteristic of the *lymphoma* measurements. Whilst those most important for being indicative of the *metastatic* measurements include: protein backbone (934-940 cm^{-1}), lipid and collagen (1296-1347 cm^{-1} and 1729-1799 cm^{-1}), protein and phospholipid (1435-1473) except in component 3, tryptophan (1538-1542 cm^{-1} and 1549-1555 cm^{-1}) and amide I (1655-1659 cm^{-1}).

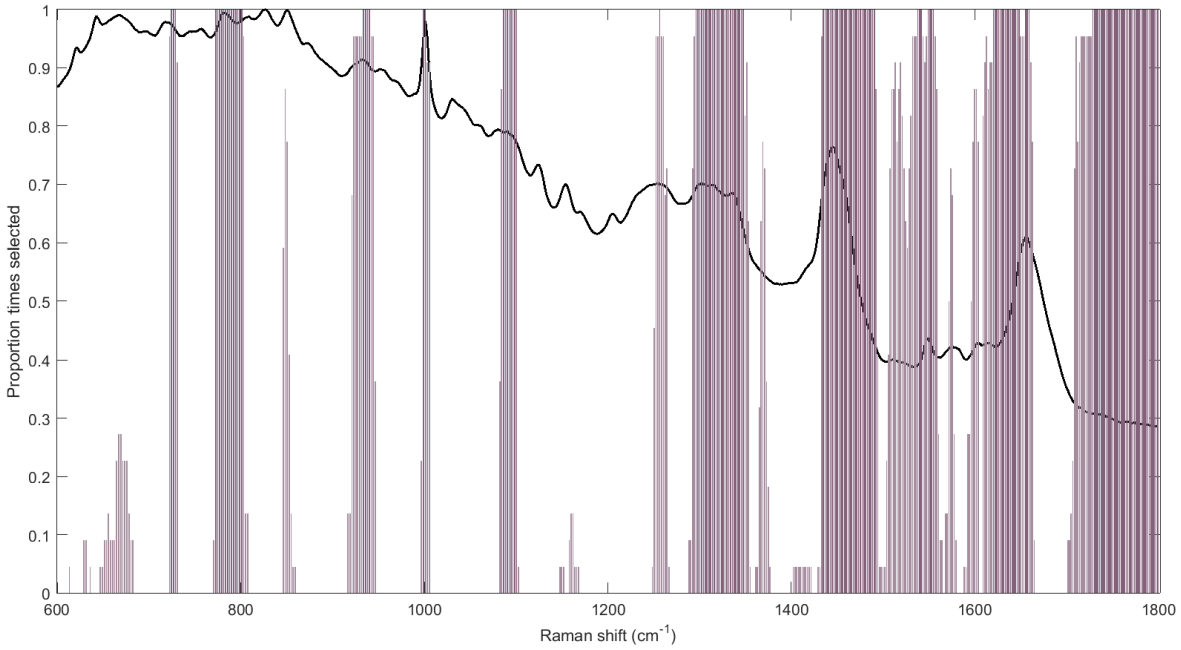

Figure 5.81- VIP plot identifying the regions of the spectrum and proportion of times selected for classification, overlaid on the dataset mean spectrum.

Table 5.41- Wavenumbers selected 100% of the time for classification and their tentative assignments for the peaks within those regions.

Bands selected 100% of times	**Tentative biochemical assignment**
724-729	DNA
773-801	Pyrimidine
934-940	Protein backbone
999-1001	Phenylalanine
1086-1101	DNA
1257	Amide III
1296-1347	Lipid, collagen, Amide III
1435-1491	Protein, phospholipid, DNA
1538-1542	Tryptophan
1549-1555	Tryptophan
1620-1648	Amino acid
1655-1659	Amide I
1729-1799	Lipid

PLSDA was carried out on just the *lymphoma* and *non-cancer* data and the scores plot is displayed in Figure 5.82. The clustering shows that the measurements group together with their own, with the exception of a few outliers. The outliers appear more frequent within *non-cancer* and may be so, due to more measurements representing this group. Overall the *lymphoma* scores are generally more negative and the *non-cancer* scores more positive across the three components. In component three the *non-cancer* scores are distributed across the axis more than in the first two components, however the majority are still positive.

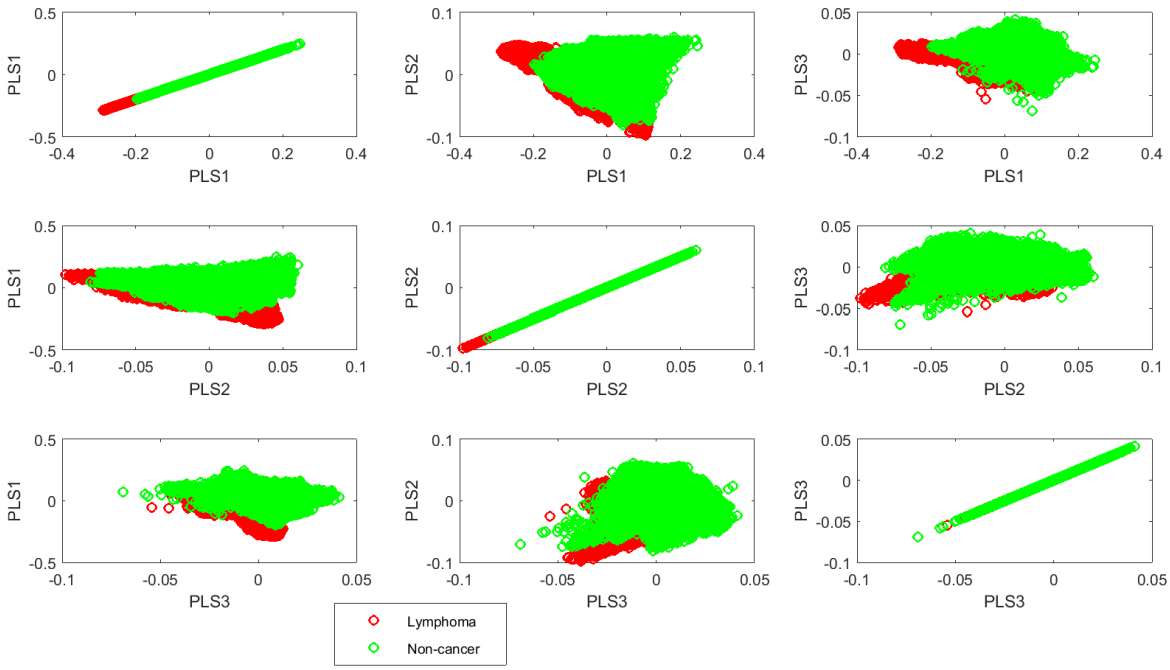

Figure 5.82- PLSDA scores plot of lymphoma (332013) and non-cancer (53879) measurements.

Cross validation was carried out to assess the classification capability of the model, the confusion matrix is presented in Table 5.42. A sensitivity and specificity of 64% and 86% for the identification of lymphoma samples from *non-cancer* lymph nodes was calculated. There appears to be a better ability of classifying the *non-cancer* samples correctly than the *lymphoma*, and again the unbalance between the group sizes may contribute to this.

Table 5.42- Confusion matrix for the cross-validation of the PLSDA model for classification of lymphoma and non-cancer measurements using a majority wins approach.

		Actual group	
		Lymphoma	**Non-cancer**
Predicted group	**Lymphoma**	9	5
	Non-cancer	5	32
	Total Samples	14	37
	Sensitivity	0.6429	0.8649
	Specificity	0.8649	0.6429

The spectra of the first three components are displayed in Figure 5.83. The first component is obviously influenced by the difference in spectral background between the groups and do not have as many spectral peaks as the other two components. Table 5.43 lists the main peaks from the PLS weights with their tentative assignments. Peaks in the negative direction, and thus more frequently observed in *lymphoma* measurements are those assigned as amino acids (656, 694, 705, 1594 and 1605 cm^{-1}), lipids and protein (1061, 1129, 1298, 1437, 1455 cm^{-1}), collagen (872 and 1298 cm^{-1}), protein (1410, 1429 and 1437 cm^{-1}) and amide I (1655 and 1659 cm^{-1}). Peaks in the positive direction, and thus more associated with *non-cancer* measurements are those assigned as DNA and nucleic acids (727, 746, 781, 1489 and 1624 cm^{-1}), carotenoid (1158 and 1521 cm^{-1}), phenylalanine (623 and 1001 cm^{-1}), tryptophan (1371, 1551, 1562 cm^{-1}), tyrosine (642 and 853 cm^{-1}), amide III (1253 and 1338). In component 3 phospholipid and protein peaks are seen in the positive direction at 1448 and 1460 cm^{-1}. There is also a putative glycogen peak at 1047 cm^{-1}, and a possible 1084 cm^{-1} peak further supports the assumed assignment. A hydroxyapatite and cholesterol peak is also present in the positive direction in component 3 at 953 cm^{-1}.

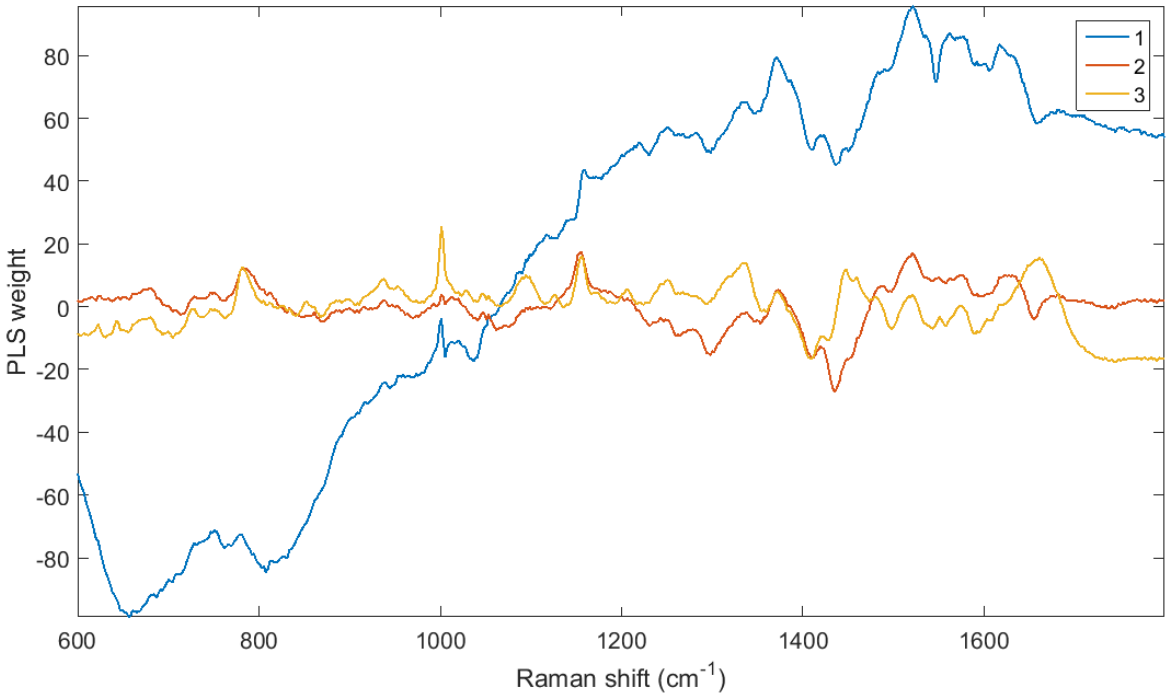

Figure 5.83-Spectra of the PLS weights of the first three components.

Table 5.43- Main peaks and their tentative assignments from the PLS weights.

PLS component 1				PLS component 2				PLS component 3			
-ve		+ve		-ve		+ve		-ve		+ve	
656	Amino acid	1001	Phe	872	Collagen	680	Amino acid	694	Cystine	623	Phe
807	Phosphodiester	1117	Lipid, protein	1061	Lipid	727	DNA	705	Cystine	642	Tyr
1129	Lipid, protein	1158	Protein, carotenoid	1261	Amide III	746	DNA	872	Collagen	680	Amino acid
1298	Lipid, collagen	1253	Amide III	1298	Lipid, collagen	781	Pyrimidine base	1061	Lipid	727	DNA
1410	Protein	1338	Amide III	1408	Protein	938	Protein backbone	1298	Lipid, collagen	746	DNA
1437	Phospholipid, protein	1371	Tryp	1455	Phospholipid, protein	1001	Phe	1408	Protein	781	Pyrimidine base
1547	Tryp	1387	Protein	1655	Amide I	1047	Glycogen	1429	Protein	853	Tyr
1594	Amino acid	1483	DNA			1156	Protein, carotenoid	1544	Tryp	938	Protein backbone
1605	Amino acid	1521	Carotenoid			1251	Amide III	1549	Tryp	953	HA cholesterol
1659	Amide I	1562	Tryp			1338	Amide III	1590	Amino acid	1001	Phe
		1617	Amino acid			1371	Tryp			1047	Glycogen
						1489	DNA			1094	DNA
						1521	Carotenoid			1127	Lipid, protein
						1551	Tryp			1156	Protein, carotenoid
						1574	Nucleic acid			1206	Protein
						1624	Amino acid			1251	Amide III
										1338	Amide III
										1371	Tryp
										1448	Phospholipid, protein
										1460	Phospholipid, protein
										1483	DNA
										1521	Carotenoid
										1551	Tryp

										1574	Nucleic acid
										1661	Amide I

VIP carried out on the cross validation of the PLSDA model suggests that the areas of most importance within the spectrum for classification are amino acids, DNA, proteins, carotenoid, amides and lipids. The plot in Figure 5.84 visualises the areas of the spectrum that contribute the most to variability and Table 5.43 lists the wavenumbers selected 100% of times along with their tentative assignments. A comparison with the weights shows that the peaks assigned as DNA (702-713 and 779-787 cm^{-1}), carotenoid (1152-1158, 1510-1528 cm^{-1}), amide III (1324-1342 cm^{-1}) and phenylalanine (629-638, 999-1003 cm^{-1}) are more representative of the *non-cancer* lymph node measurements. Whilst the peaks assigned as tyrosine (647-658 cm^{-1}), tryptophan (1552 and 1557-1564 cm^{-1}), amino acids (1587-1596 and 1617 cm^{-1}), protein (1402-1418) and amide I (1617-1678 cm^{-1}) more frequently represent lymphoma based on the scores plot.

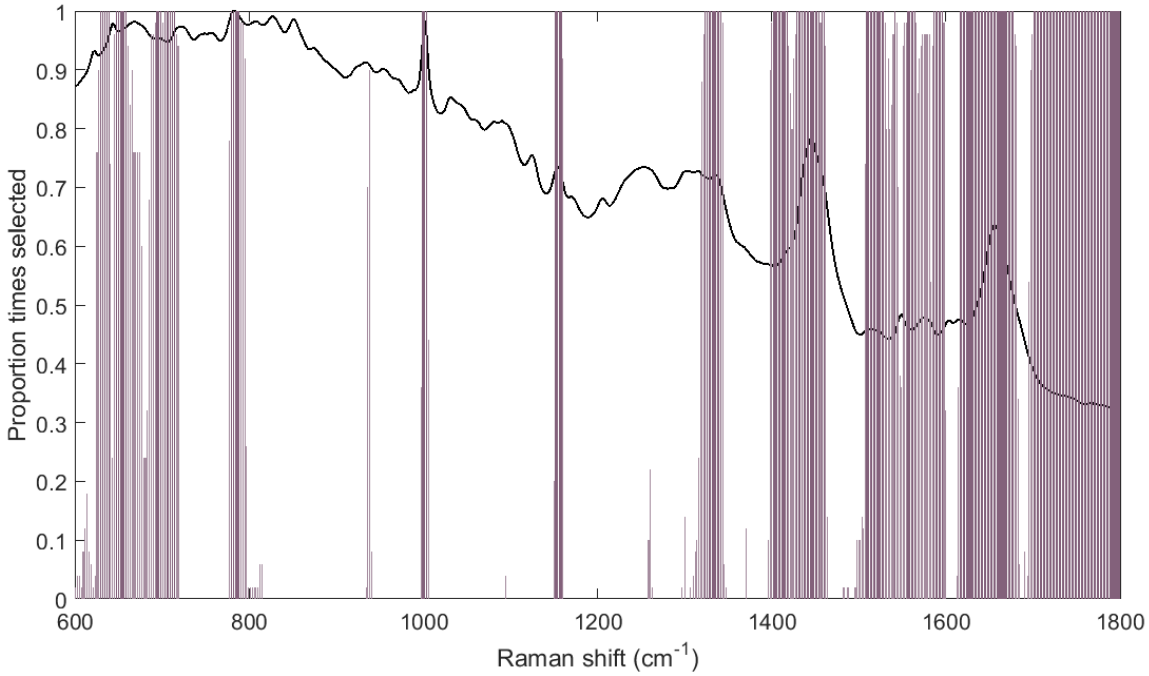

Figure 5.84- VIP bar graph showing the proportion of spectral regions selected for cross-validation, plotted over mean spectrum.

Table 5.44- Wavenumbers selected 100% of times for cross-validation with tentative assignments.

Bands selected 100% of times (cm^{-1})	**Tentative biochemical assignment**
629-638	Phenylalanine
647-658	Tyrosine, amino acid
694-697	Cystine
702-713	DNA
779-787	Pyrimidine base
999-1003	Phenylalanine
1152-1158	Protein, carotenoid
1324-1342	Amide III
1402-1418	Protein
1429-1460	Protein, Phospholipid
1510-1528	Carotenoid
1542	Tryptophan
1557-1564	Tryptophan
1587-1596	Amino acid
1617-1678	Amino acid, Amide I
1701-1799	Lipid

The PLSDA scores plot, Figure 5.85, shows the clustering of *metastatic* measurements and *non-cancer* measurements. The separation between the two groups is more distinct than in the *lymphoma* and *non-cancer* model, though not entirely delineated. Once more, outliers are present in both groups but with a higher frequency and at a greater degree in the *non-cancer* data than the *metastatic*. The *non-cancer* data is more typically positive across the components whilst the *metastatic* is more often negative across the components.

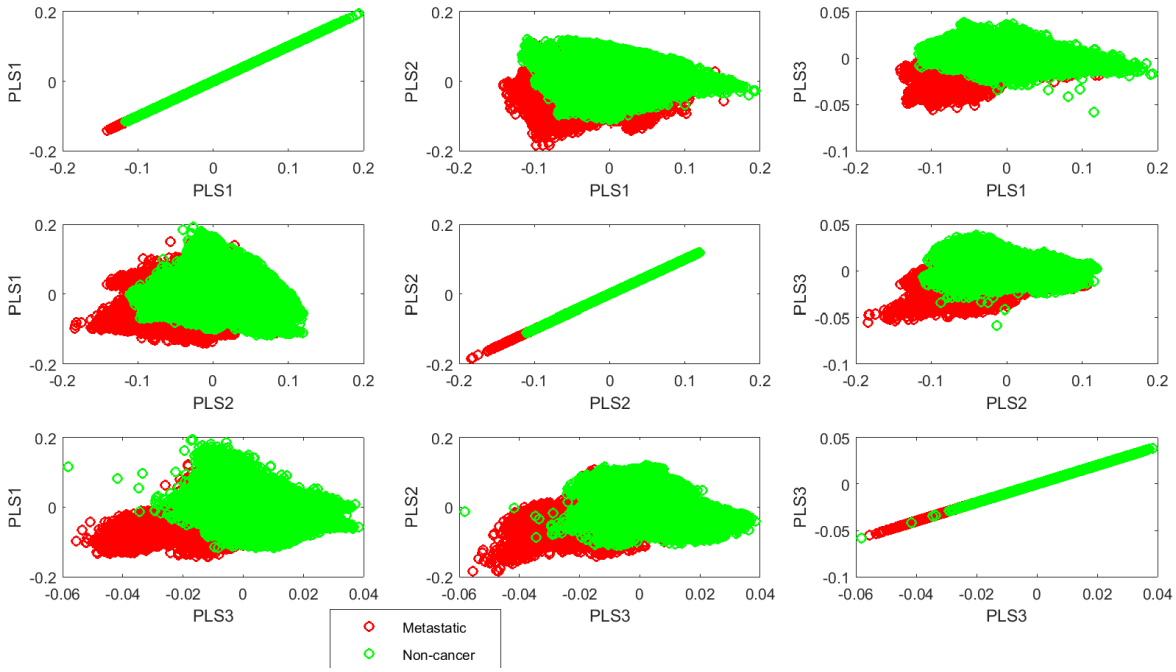

Figure 5.85- PLSDA scores plot for the Raman mapped metastatic (12,120 spectra) and non-cancer (53,879 spectra) lymph node data.

Cross-validating the model produced a sensitivity of 75% and specificity of 97% for the identification of *metastatic* lesions in lymph nodes from *non-cancer* lymph nodes, with only one *non-cancer* lymph node being mistaken for *metastatic* (Table 5.45). From these results it is apparent that the *metastatic* node measurements and those of the *non-cancer* lymph nodes are expectedly more different to one another than the *lymphoma* and *non-cancer* measurements, most probably as a result of the difference in cell types.

Table 5.45- Confusion matrix for the cross-validation of the PLSDA model using a majority wins approach.

		Actual group	
		Metastatic	**Non-cancer**
Predicted group	**Metastatic**	6	1
	Non-cancer	2	36
	Total Samples	8	37
	Sensitivity	0.7500	0.9730
	Specificity	0.9730	0.7500

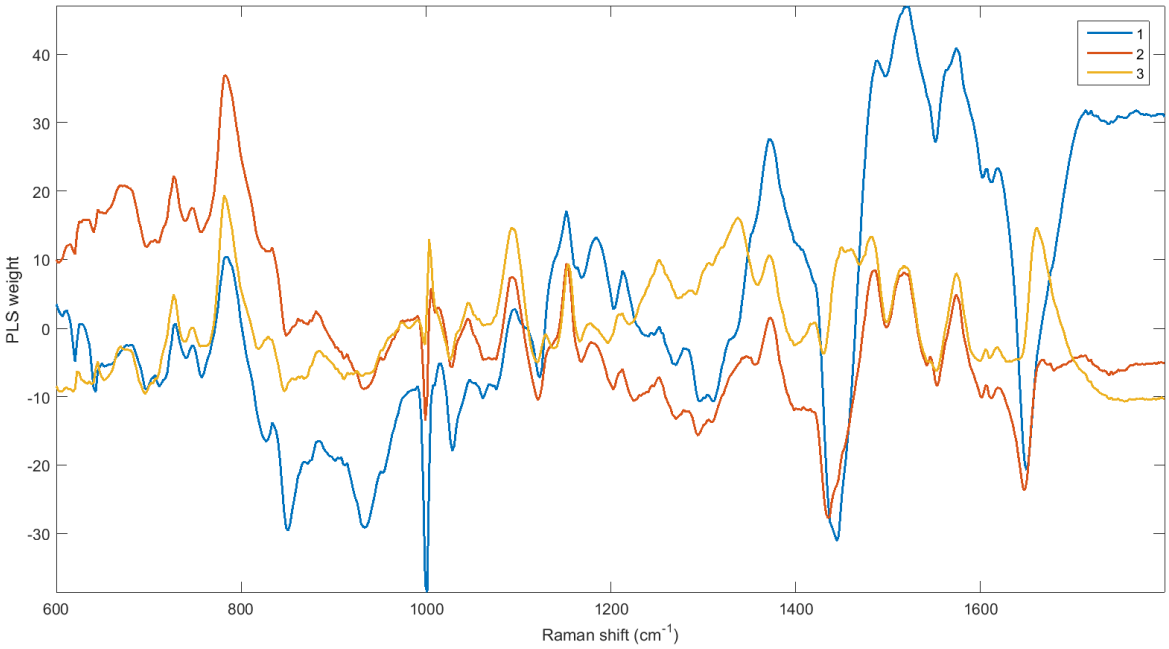

Figure 5.86- Spectra of PLS weights from the first three components.

The weight spectra in Figure 5.86 were evaluated to gain a better understanding of the spectral changes and hence biochemical contributions that lead to differences between the two groups. Table 5.46 lists the main peaks from the weights along with the tentative biochemical assignments to see what molecules are contributing to the variation across the pathological groups. Those consistently observed in the positive direction across all the components, thus more indicative of the measurements from the *non-cancer* lymph node tissue include peaks assigned as DNA and nucleic acid (729, 749, 781, 833, 1092, 1487 and 1574 cm^{-1}), carotenoid (1152 and 1517 cm^{-1}), amide III (1253 and 1338 cm^{-1}), glycogen (1045 cm^{-1}), tyrosine (1180 cm^{-1}) and tryptophan (1373 cm^{-1}). Collagen assigned peaks at 883 and 1278 cm^{-1} were observed in the positive direction although also present in the negative weights, which is more suggestive of *metastatic* measurements, at 872 and 1270 cm^{-1}.

Table 5.46- Main peaks from the weights with their tentative biochemical assignments.

PLS component 1				PLS component 2				PLS component 3			
-ve		+ve		-ve		+ve		-ve		+ve	
620	Phe	729	DNA	620	Phe	644.6	Tyr	698	Cystine	622	Phe
642	Tyr	749	DNA	640	Tyr	669.1	Amino acid	709	Amino acid	644	Try
695	Cystine	781	Pyrimidine base	698	Cystine	727	DNA	851	Tyr	727	DNA
757	Tryp	833	Nucleic acid	934	Protein backbone	746	DNA	1001	Phe	746	DNA
827	Nucleic acid	1092	DNA	1001	Phe	784	Pyrimidine base	1028	Phe	781	Pyrimidine base
851	Tyr	1152	Protein, carotenoid	1028	Phe	833	Nucleic acid	1121	Lipid, protein	833	Nucleic acid
872	Collagen	1180	Tyr	1121	Lipid, protein	861	Tyr	1398	Protein	883	Collagen
934	Protein backbone	1212	Protein	1270	Collagen, Amide III	883	Collagen	1431	Phospholipid, protein	1003	Phe
955	Hydroxyapatite, cholesterol	1253	Amide III	1294	Lipid, collagen	1003	Phe	1498	DNA	1045	Glycogen
1001	Phe	1278	Collagen	1355	Tryp	1045	Glycogen	1542	Tryp	1092	DNA
1029	Phe	1371	Tryp	1435	Phospholipid, protein	1092	DNA	1551	Tryp	1129	Lipid, protein
1061	Lipid	1487	DNA	1492	DNA	1152	Protein, carotenoid	1605	Amino acid	1152	Protein, carotenoid
1076	Lipid	1521	Carotenoid	1542	Tryp	1180	Tyr	1618	Amino acid	1180	Tyr
1123	Lipid, protein	1574	Nucleic acid	1551	Tryp	1212	Protein			1212	Protein
1270	Collagen, amide III			1605	Amino acid	1253	Amide III			1253	Amide III
1294	Lipid, collagen			1618	Amino acid	1280	Collagen			1280	Collagen
1310	Lipid, collagen			1648	Amide I	1346	Amide III			1338	Amide III
1445	Phospholipid, protein					1373	Tryp			1373	Tryp
1551	Tryp					1485	DNA			1421	Protein
1602	Amino acid					1517	Carotenoid			1450	Phospholipid, protein
1611	Amino acid					1574	Nucleic acid			1483	DNA
1650	Amide I									1517	Carotenoid

										1661	Amide I

VIP on the PLSDA model (Figure 5.87) shows that the bands selected 100% of times for cross-validation include those assigned to DNA and nucleic acids, amino acids, proteins, carotenoids, collagen, amides and lipids. A list of these bands is displayed in Table 5.47. Through comparison to the weights, important peaks that are characteristic of *non-cancer* LN measurements constitute those assigned as DNA and nucleic acid (784, 1092, 1485 and 1574 cm^{-1}), protein and carotenoid (1152 and 1517 cm^{-1}), amide III (1346 cm^{-1}). Peaks characteristic of *metastatic* measurements and used 100% of times for cross-validation include tyrosine (851 cm^{-1}), phenylalanine (1001 cm^{-1}), collagen (1310 cm^{-1}), protein and phospholipid (1445 cm^{-1}), tryptophan (1551 cm^{-1}) and amide I (1650 cm^{-1}).

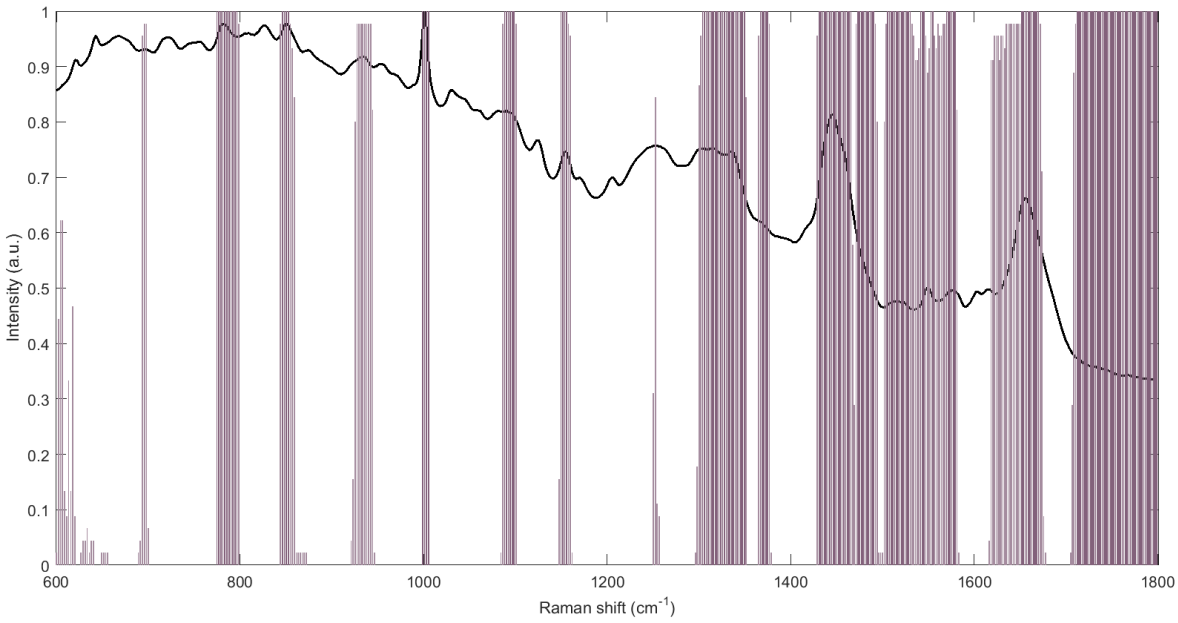

Figure 5.87- VIP bar graph plotted over mean spectrum.

Table 5.47- VIP regions selected 100% of times for cross-validation with their tentative assignments.

Bands selected 100% of times	**Tentative biochemical assignment**
774-797	Pyrimidine
846-853	Tyrosine
999-1005	Phenylalanine
1088-1099	DNA
1150-1156	Protein/carotenoid
1304-1349	Collagen, amide III
1367-1375	Tryptophan
1431-1466	Protein, phospholipid
1473-1491	DNA
1506-1530	Carotenoid
1542-1545	Tryptophan
1553-1555	Tryptophan
1570-1579	Nucleic acid
1652-1670	Amide I
1712-1799	Lipid

The mean spectra of the metastatic head and neck lymph nodes, *SCC*, *AC* and *PC*, are displayed in Figure 5.88. There are no obvious visual differences between the three averaged spectra of the metastatic tissue measurements although they originate from different cell types.

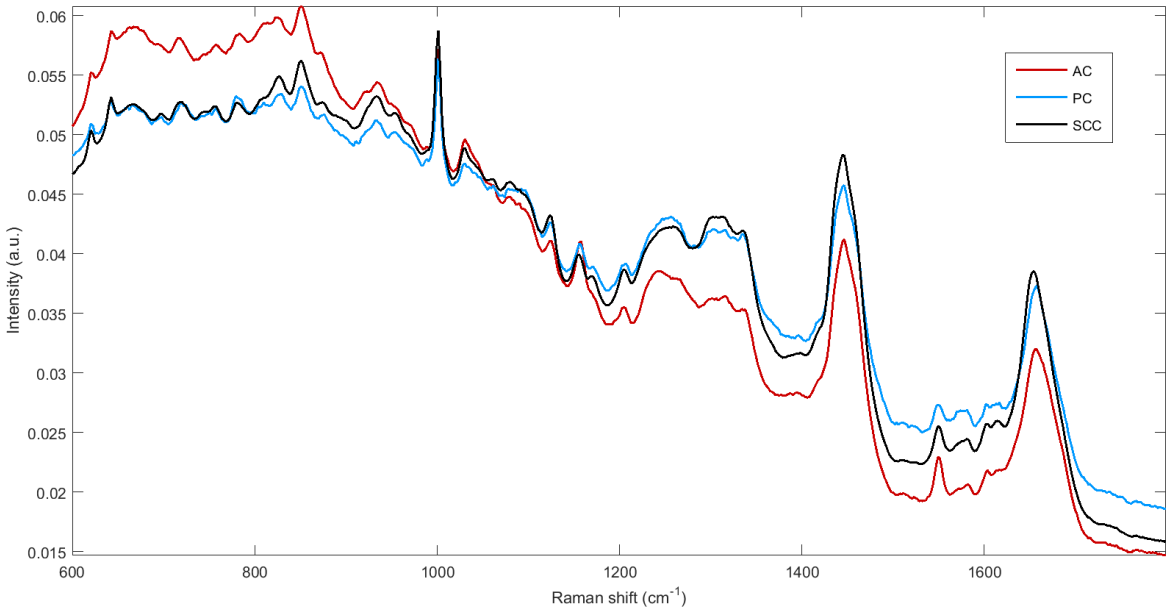

Figure 5.88- Mean spectra of measurements from AC (1076 spectra), PC (714 spectra) and SCC (10330 spectra) metastases in head and neck lymph nodes.

Only one *AC* and *PC* sample were collected for this study, as they are not as common in head and neck lymph nodes as *SCC*, and therefore leave-one-sample-out cross-validation could not be conducted. However, PLSDA was carried out and its scores plot is displayed in Figure 5.89 and shows the distribution of the data across the components. *AC* and *SCC* appear to be the most clearly different groups, whilst the *PC* data appears to be more similar to the *SCC* group as oppose to the *AC*. As the *AC* and *PC* measurements are taken from only one node each, separation between the groups may be as a result from biochemical differences between the lymph nodes as oppose to biochemical alterations between pathologies. However, the data within the *AC* and *PC* groups do not cluster as tightly as expected, considering they are from single samples. Typically the *AC* scores are more negative across the three components whilst the *SCC* are generally positive and the *PC* scores more positive in component one and negative in components two and three.

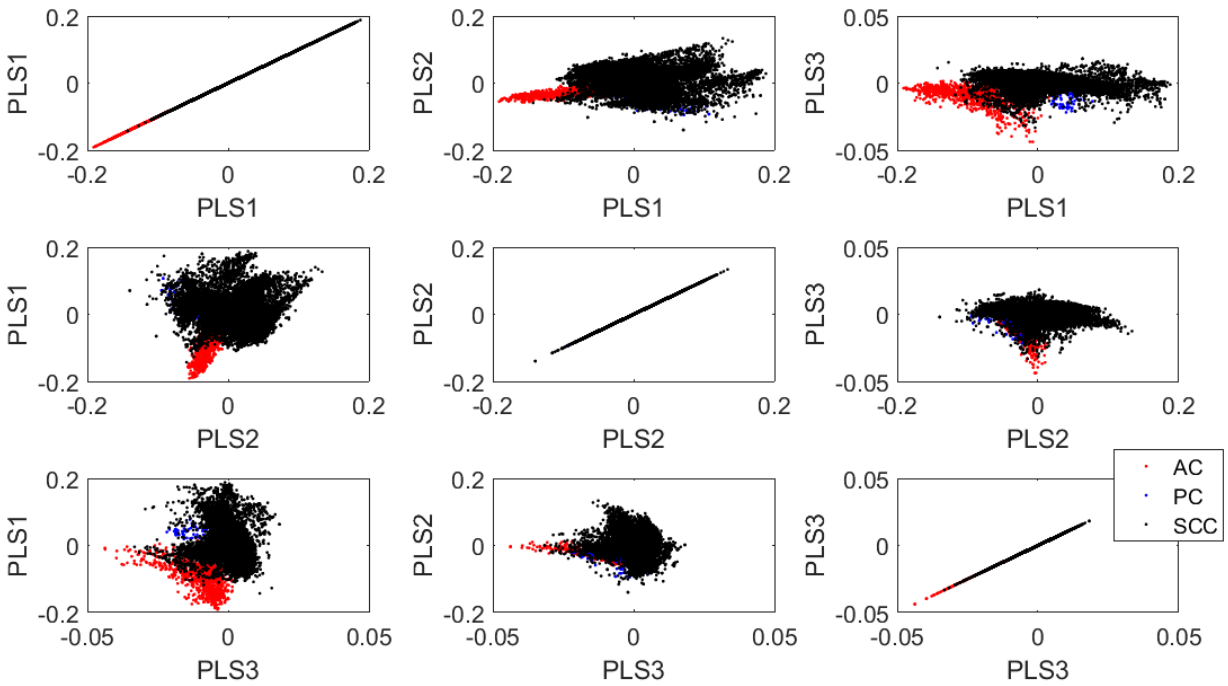

Figure 5.89- PLS scores plots of AC, PC and SCC met data from the first three components

The weights for the components, displayed in Figure 5.90, show the main peaks contributing to the variation between the groups. Their tentative biochemical assignments are listed in Table 5.48. The main spectral differences lies in the DNA and nucleic acid region of 600-900 cm^{-1} where there are many peak variations. There is also a split peak at 999/1003 cm^{-1} suggesting a shift in the phenylalanine peak between the groups. Amide I (1648 cm^{-1}) and III (1260 cm^{-1}) also appear to contribute to variation between groups along with proteins in general (1074, 1121, 1158, 1200, 1218 and 1433 cm^{-1}).

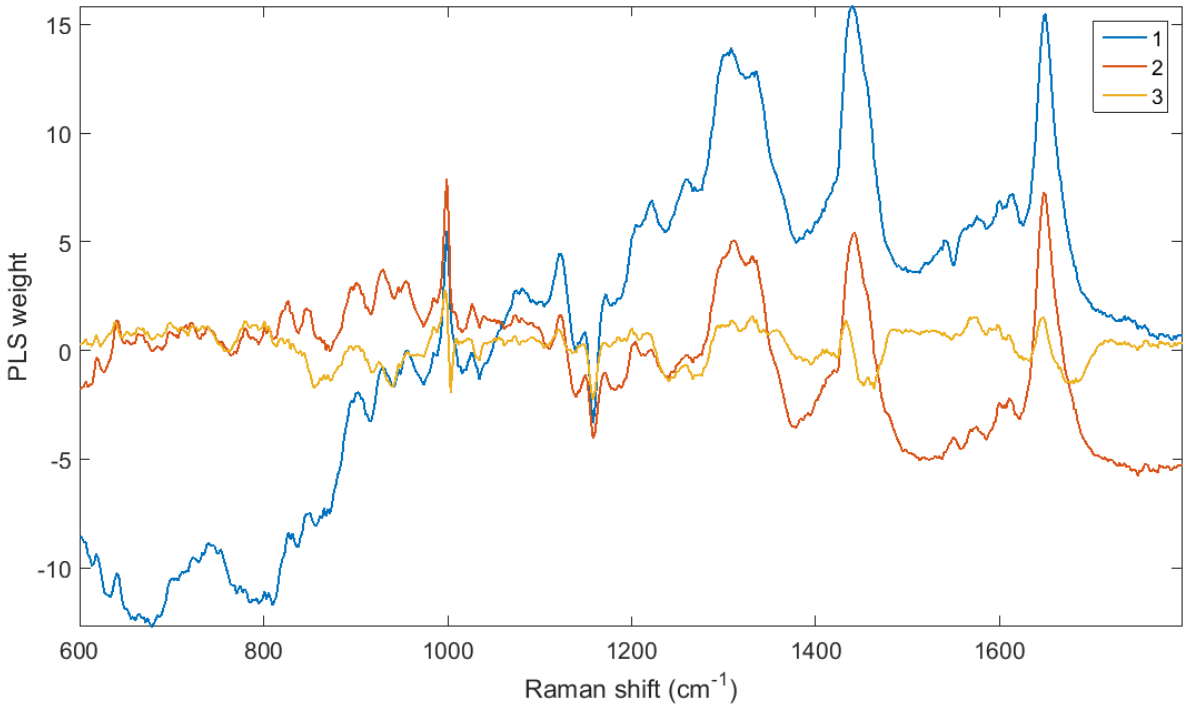

Figure 5.90- Spectra of the PLS weights of the first three components.

Table 5.48- The main peaks from the component weights and their tentative biochemical assignments.

PLS component 1				PLS component 2				PLS component 3			
-ve		+ve		-ve		+ve		-ve		+ve	
628	Phe	827	Nucleic acid	872	Collagen	618	Phe	855	Tyr	618	Phe
678	Amino acid	851	Tyr	973	Collagen	638	Amino acid	872	Collagen	638	Amino acid
770	Pyrimidine	902	Adenine	1158	Protein, carotenoid	665	Amino acid	940	Protein	698	Cystine
810	Phosphodiester	930	Protein			698	Cystine	1003	Phe	779	Pyrimidine base
1034	Phe	955	HA, cholesterol			722	DNA	1034	Phe	900	Adenine
1158	Protein, carotenoid	999	Phe			740	DNA	1158	Protein, carotenoid	999	Phe
1551	Tryp	1026	Pe			779	Pyrimidine base	1452	Phospholipid, protein	1121	Lipid, protein
		1082	Protein, lipid			801	Phosphodiester			1200	Protein, amino acid
		1121	Lipid, protein			827	Nucleic acid			1218	Protein
		1150	Protein, carotenoid			846	Tyr			1259	Amide III

		1172	Tyr			900	Adenine			1308	Lipid, collagen
		1222	Protein			930	Protein			1332	Amide III
		1260	Amide III			955	HA, cholesterol			1433	Phospholipid, protein
		1308	Lipid, collagen			999	Phe			1572	Nucleic acid
		1336	Amide III			1026	Phe			1598	Amino acid
		1441	Phospholipid, protein			1074	Protein, lipid			1611	Amino acid
		1576	Nucleic acid			1121	Protein, lipid			1648	Amide I
		1600	Amino acid			1200	Protein				
		1615	Amino acid			1218	Protein				
		1648	Amide I			1312	Collagen				
						1332	Amide III				
						1443	Phospholipid, protein				
						1549	Tryp				
						1576	Nucleic acid				
						1600	Amino acid				
						1611	Amino acid				
						1648	Amide I				

The average spectra of the *HL* and *NHL* acquisitions are plotted in Figure 5.91. The main visual difference between the *HL* and *NHL* spectra appear to be within in the 770-890 cm^{-1} wavenumber region. However, the background may play a part in this and disguise the peaks that are clearly present in the *NHL* spectrum. The 1152 cm^{-1} peak assigned as protein and carotenoid also appears to be more intense in the *HL* spectrum than the *NHL*.

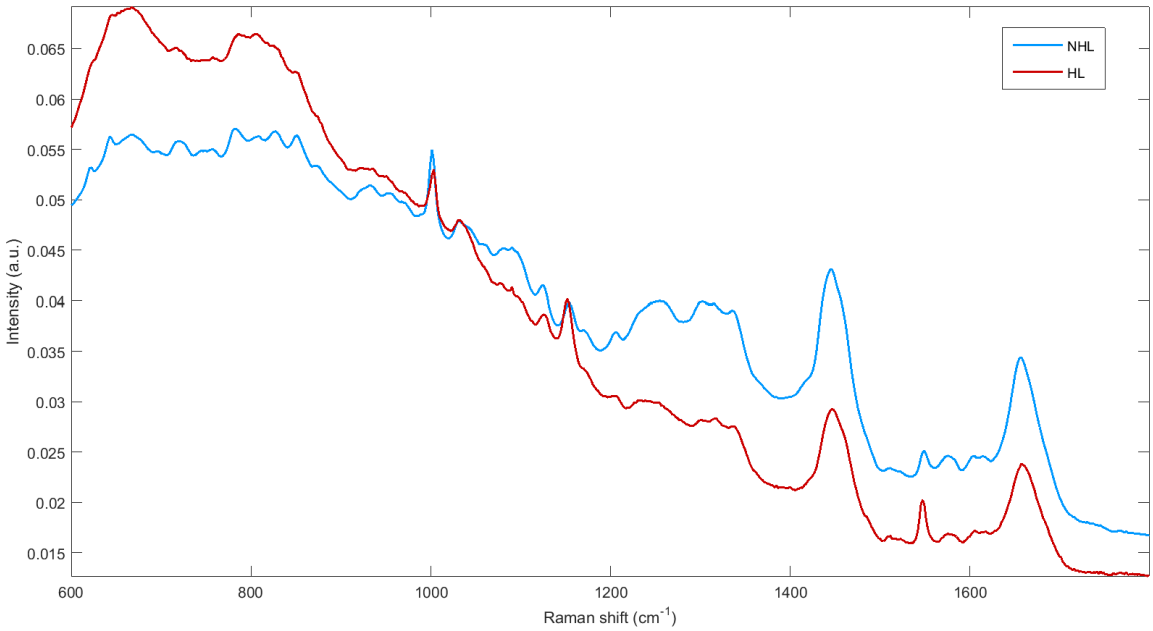

Figure 5.91- Mean spectra of HL (2690 spectra) and NHL (30511 spectra) measurements.

The PLSDA scores plot, displayed in Figure 5.92, shows the *HL* measurements to be very negative in the three components. The *NHL* measurements are much more dispersed in the three components although in general more positive.

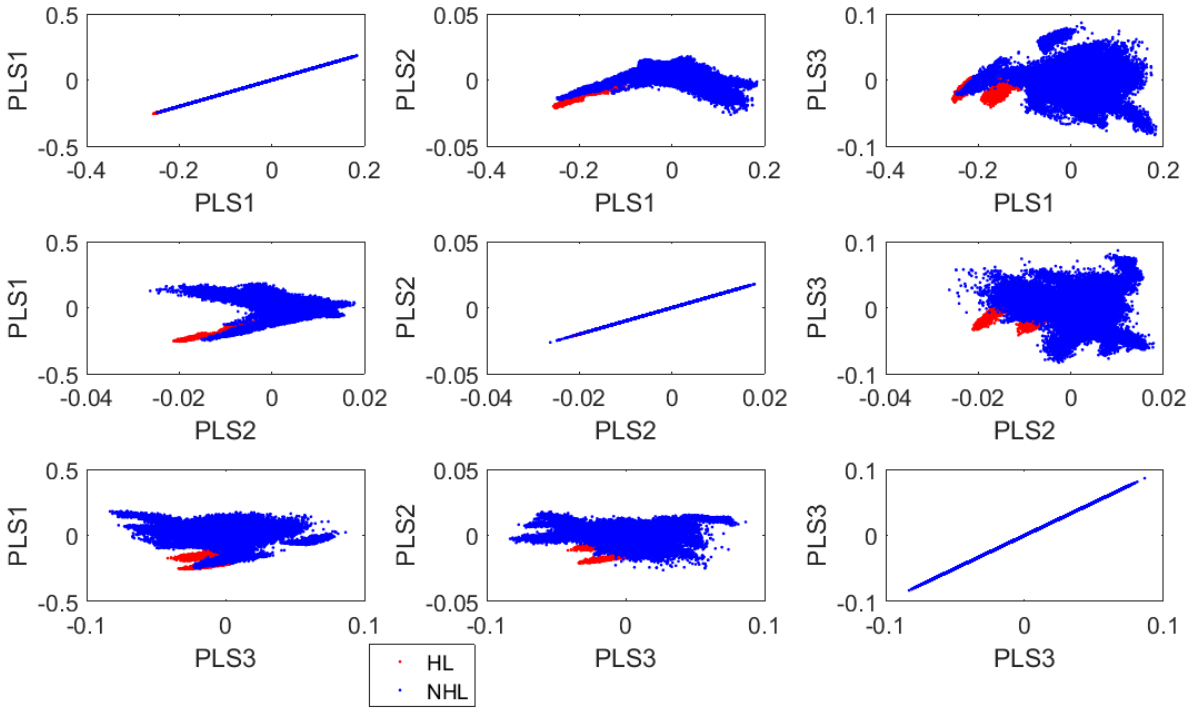

Figure 5.92- PLSDA scored plot of the HL and NHL data from the first three components.

The weights from the components, Figure 5.93, show the background as a heavy influence in the first component. Table 5.49 lists the main peaks from the weights and their tentative biochemical assignments. Peaks within the DNA region around 600-900 cm^{-1} do not appear to be contributing to variation, except in component 1, but the peaks in this region could be affected by the background issue. The phenylalanine assigned peak at 1003 is seen to be contributing to the variation between the groups, along with amide III (1261, 1340 cm^{-1}) and I (1655 cm^{-1}). Peaks assigned as amino acid (1367, 1545, 1602 and 1615 cm^{-1}) are also present in the component weights and also lipids and protein (1125, 1204, 1308, 1414 and 1441 cm^{-1}). The majority of the peaks contributing to the variation appear to be present in component 1 however, as component 2 and 3 do not seem to have as many spectral peaks present.

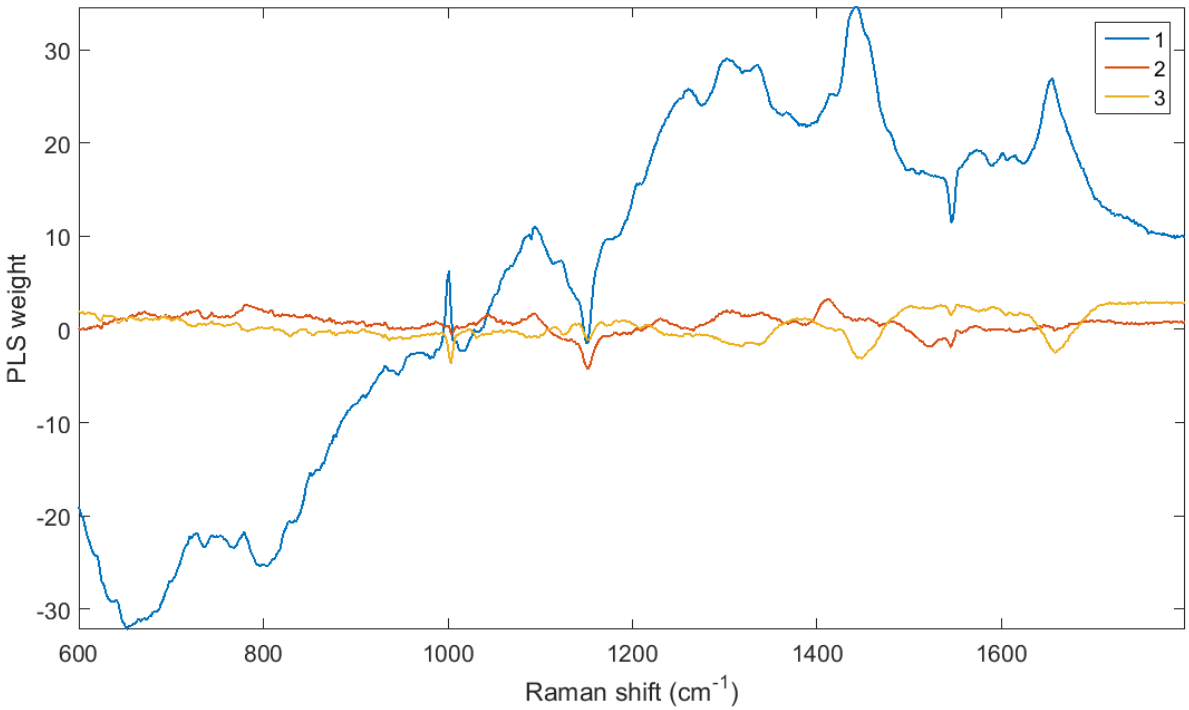

Figure 5.93-Spectra of the PLS weights of the first three components.

Table 5.49- The main peaks from the component weights and their tentative biochemical assignments.

PLS component 1				PLS component 2				PLS component 3			
-ve		+ve		-ve		+ve		-ve		+ve	
805	Phosphodiester	729	DNA	1003	Phe	781	Pyrimidine	1003	Phe		
1152	Protein, carotenoid	749	DNA	1152	Protein, carotenoid	1043	Glycogen	1127	Lipid, protein		
1545	Tryp	780	Pyrimidine base	1521	Carotenoid	1094	DNA	1152	Protein, carotenoid		
		1001	Phe	1545	Tryp	1231	Protein, amide III	1445	Phospholipid, protein		
		1094	DNA			1308	Lipid, collagen	1659	Amide I		
		1125	Lipid, protein			1340	Amide III				
		1204	Protein			1414	Phospholipids, protein				
		1261	Amide III								
		1302	Lipid, collagen								
		1336	Amide III								
		1367	Tryp								
		1414	Protein, lipid								
		1441	Phospholipid, protein								
		1602	Amino acid								
		1615	Amino acid								
		1655	Amide I								

Discussion

Along with metastatic tumours, lymphomas are also frequently diagnosed in the lymph nodes of the head and neck. Often discovered due to their visual enlargement and hardening, prior to any other symptoms. However, lymph nodes are also known to become inflamed after infection as part of the immune response. Therefore it is vital that inflammation as a result of a benign condition is not mistaken for cancer, and more importantly that a malignancy is not incorrectly classified as a reactive node. It is also essential for treatment to correctly classify the type of cancer (the treatment for NHL is different to HL), and whether it is secondary or primary to further search for the cancer origin and subsequent removal if required. Furthermore, lymph nodes are essential organs for aiding immunity and tissue drainage and removal of them can cause a host of problems for a patient. Therefore *in vivo* diagnosis of lymph nodes would be highly beneficial for both patients and clinical practices, reducing the need for surgery and time for diagnosis.

A decreased peak intensity at 1515 cm^{-1}, assigned as carotenoid in the measurements of the *metastases* and *lymphoma* samples when compared with the benign groups aligns with the findings from the LN studies from other locations. A decrease of this peak in the *lymphoma* measurements as well as the *metastatic*, indicates that the decrease in carotenoid is due to the tissue becoming cancerous as opposed to a difference in cell type between the groups.

The visual similarity between the two benign groups, labelled *non-cancer* and *other*, suggest that spectral differences are more prominent between malignant and benign measurements than just reactive diseases measured in the lymph nodes. Only three samples were in the *other* group, however and a larger investigation into diseased nodes would be beneficial to learn the differences between cancer tissues from other diseased tissues.

Surprisingly, a better sensitivity and specificity of 82% and 90% were calculated when including the *other* group in the PLSDA and classifying it together with *non-cancer* as three *other* nodes were classified correctly as *non-cancer*. Also, one more *cancer* sample was classified correctly in this model than

when the *other* group were excluded, suggesting the presence of this group may have helped the classification of *cancer* samples.

Using an ROC curve would be helpful for clinical settings, especially when implementing new technologies, where 100% sensitivity would be desired even at the expense of specificity. If this technique was used *in vivo*, a 100% sensitivity could be implemented to initially assess the lymph nodes in situ and still prevent a number of biopsies and hence save money and time for both surgeons and pathologists. Those that were biopsied could then be assessed by pathologists in the usual manner and the reduced number of biopsies could enable more time being spent on and more pathologists examining difficult sections. The PLSDA model of *cancer* and *non-cancer* could achieve a classification with 100% sensitivity but with a specificity of just 28%. Even still, this would have clinical benefit by reducing unnecessary surgery for many patients whilst still ensuring cancer lesions are not missed.

The main differences between the *cancer* and *non-cancer* samples appear to be as a result of spectral peaks assigned as Lipids, DNA, amino acids and proteins. Analysis of the PLS scores plot and weights suggest that the *cancer* samples appear to have more intense lipid, protein (especially collagen) and amide I assigned peaks, and a decrease in DNA, amide III, glycogen and carotenoid assigned peaks.

Although DNA content is supposed to be consistent across diploid cell types at around 7 billion base pairs stored in the nucleus, Gillooly *et al*.[268] observed a difference in DNA concentration between different cell types related to their cell size. They observed a positive correlation between cell volume and DNA content. Megakaryocytes were observed to contain both the highest level of DNA and have the highest cell volume. Whilst the cell volume of lymphocytes is slightly higher than many cells types, its DNA concentration is standard and therefore DNA quantity of individual cells does not seem to be contributing to a higher DNA quantity measured in the *non-cancer* samples. As mentioned, Lloyd *et al*.[58] explained that lymph nodes contain densely packed lymphocytes that have been shown to exhibit much stronger DNA signal than epithelial cells. Therefore, it is more likely as a result of a higher quantity of cells and thus overall

DNA quantity in the tissue and its stronger signals that are solely contributing to this higher detected level in the non-cancer lymphocyte cells. As the *metastatic* and *lymphoma* samples are classed as one group, lower DNA signals in the epithelial cells may affect this difference in the DNA peaks between the *non-cancer* and the *cancer* measurements. Additionally, as Lloyd et al.[58] also mentioned, metastatic invasion can also reduce the nucleic acid density. Although most of the measured DNA peaks are more characteristic of the positive scores, and thus primarily *non-cancer* data, a peak at ~1490 cm^{-1} was observed to be more indicative of the negative scores and thus more often the *cancer* data.

The tumour microenvironment has a major influence in tumour development. The extracellular matrix (ECM) is a vital component of this and not only provides the binding scaffold but also influences biochemical and physical signalling. Collagen is the most abundant molecule in the ECM and its role in tumour development has been well documented.[269] Accounting for approximately one third of proteins, collagen is the most common, existing in many different forms. A higher presence of collagen is usually expected in tumour tissue than normal. Sok *et al*.[270] found collagen type XI at elevated levels in patients with head and neck SCC and were able to demonstrate that it contributes to proliferation, migration and invasion through real-time polymerase-chain-reaction (RT-PCR). Type XI collagen is reported to help stabilise collagen II but its exact role remains unclear.

A peak at 1045 cm^{-1} was assigned as glycogen. A peak at 1082 cm^{-1}, although not obvious in the weights is seen in the spectrum. This peak further indicates this is indicative of glycogen. The Raman spectrum of glycogen also have peaks at 854, 939, 1131, 1257, 1336, 1381 and a strong peak at 459 cm^{-1}. However, this area of the spectrum was not measured and therefore could not be used to help confirm or reject a glycogen assignment. This glycogen peak was found to be more indicative of *non-cancer* measurements than *cancer* in this study. Kamemento *et al*.[216] also observed a reduction of glycogen peaks in cancer tissue and noted that Lyng *et al*.[137] reported that glycogen is acquired during cell maturation and may disappear when there is a loss of differentiation such as in neoplasia.

Lipids in adipose tissue are primarily in the form of triacylglycerol (TAG) and their main role is to accumulate energy and aid metabolism. Czamara *et al*.[271] stated that Raman spectroscopy is a useful technique for lipid analysis, especially as the spectral profile is significantly different depending on saturation, state, isomerism and form. They explain that the most characteristic features are due to hydrocarbon chains, and observed in the regions of 1050-1200, 1250-1300 cm^{-1} and 1400-1500 due to C-C, CH_3 and CH_2 bonds respectively. The most important marker for TAGs was the C=O vibration between 1720-1750 cm^{-1}. This may explain why there are visible shifts between the lipid peaks measured from *cancer* and *non-cancer* LN tissue and why they are always present in the weights and the VIP plots, contributing to the variation. Manabe *et al*.[272] stated that tumour mass is controlled by the balance of tumour cell growth and apoptosis of tumour cells. They reported that mature adipocytes did not influence apoptosis however accelerate growth of tumour cells. Also that a large concentration of adipose tissue is associated with high cellular proliferation and poor prognosis in breast cancer patients. They suggested that the reason for this could be as a result of increased oestrogen produced by the adipose tissue or as a result of a higher supply of energy stored in and sourced from the adipose tissue.

Gniadecka *et al*.[187] observed a major decrease in the amide I band in malignant melanoma when compared with normal skin tissue. They also reported shifts in the amide bands between normal and melanoma. In their study, the amide I band appeared at a decreased wavenumber shift whilst amide III is observed at a higher shift in the melanoma samples than when compared to normal tissue. Whilst the amide I peak is recognised in the weights as more indicative of *cancer*, the mean spectra show that in the individual *lymphoma* spectrum there appears to be a decrease in amide I. Lo *et al*.[273] observed the decrease in both amide I and III bands as the keratinocytes measured matured whilst the phenylalanine peak remained constant. This could affect the amide content in lymph nodes where the infiltration of SCC metastases have resulted in keratinocytes. The presence of the amide I peak in the weights could therefore be a result of a higher concentration of immature keratinocytes in the *metastatic* samples or perhaps a result of a peak shift to 1652 cm^{-1} from 1656 cm^{-1}.

The negative correlation between the consumption of foods high in carotene and cancer diagnosis has long been reported and the anti-carcinogenic effects of carotenoids have been researched.[274] Carotenoids' ability to inhibit cancer have been attributed to its pro-vitamin A activity, the anti-oxidant properties of the molecules or a combination of both.[275] Huang *et al*.[199] also observed the decrease of carotenoid attributed peaks in tumour tissue of bronchial tissue.

The VIP carried out on the data showed the areas of the spectrum that most contributed to variation through the proportion of times it was selected for classification. DNA, amino acids, lipids and proteins were the main biochemicals assigned to the spectral areas, selected 100% of times for classification and thus are important for identifying cancer from non-cancer measurements.

When the *cancer* and *non-cancer* groups were separated further into pathologies the classification accuracies were worse. In the four group model the sensitivities calculated from cross-validation were 50%, 25%, 33% and 92% for *lymphoma*, *metastases*, *other* and *non-cancer* and the specificities calculated were 90%, 94%, 95% and 68%. The groups with the worse sensitivities were also the two with the least samples. Low sample numbers will likely contribute to a poorer sensitivity, especially as the other groups had much larger sample-sets leading to an imbalance across the groups. Whilst the *non-cancer* classification had a high sensitivity, its specificity was the worse of the four groups. Evaluation of the confusion matrix shows that when misclassification occurs, *lymphoma* are more likely incorrectly identified as *non-cancer*. Incorrect *metastatic* and *other* classifications are distributed fairly evenly throughout the various groups and *non-cancer*, although only three nodes misclassified, are only incorrectly identified as *lymphoma* or *other*. The misclassifications suggest that the *lymphoma* and *non-cancer* groups are more similar to each other than to the *metastatic*. Although the *lymphoma* samples contain malignant lymphocytes, they may be more similar to the benign lymphocytes than they are to the malignant epithelial cells in the *metastatic* samples. The *non-cancer* samples also seem to be more similar to the *other* and *lymphocyte* group than to the *metastatic* which is also likely as a result of the presence of primarily lymphocyte cells as oppose to *metastatic* from other tissue types. Although the misclassified

metastatic samples are more evenly distributed they are more likely classed as *lymphoma* than either other or *non-cancer*, possibly due to the groups shared malignancy. As the *other* was a small group, of only three samples, it is likely this lead to the inability to classify any subsequent samples correctly, as only two samples were used for training the model before it was tested for validation. Another contributing factor could be as a result of the samples containing different benign diseases, and therefore they did not share generic similarities that distinguished them from the other groups in the model.

The combination of the *other* and *non-cancer* groups led to slightly better classification but still poorer for *metastatic*, maybe again as a result of a small sample set. Sensitivities of 57%, 38% and 90% were calculated for classification of *lymphoma*, *metastatic* and *other* combined with *non-cancer* and specificities of 88%, 96% and 68%, respectively. When misclassified, *lymphoma* was more likely to be identified as *non-cancer* than the *metastatic,* and *metastatic* was more likely to be identified as *lymphoma*.

When the *other* group was removed from analysis, PLSDA and cross-validation led to sensitivities of 57%, 38% and 89% and specificities of 87%, 96% and 68% for *lymphoma*, *metastatic* and *non-cancer* classification, respectively. The only difference in the confusion matrix is three less correctly identified *non-cancer* samples. Thus, suggesting that the three *other* samples that were diseased yet benign were more similar to the *non-cancer* samples when grouped with them than some of the completely non-diseased samples.

A two group model was then carried out with the *lymphoma* and *metastatic* data to identify if spectral differences could be identified. Sensitivities of 86% and 63% were calculated for classifying *lymphoma* and *metastatic* tissue respectively. More *metastatic* samples were incorrectly identified as *lymphoma* than *lymphoma* as *metastatic*. Again, the lack of *metastatic* samples compared to *lymphoma* may contribute to this. It may also be a result of three different *metastatic* groups where the tumour tissues are all from different primary sites, resulting in more difficulty in identifying a relationship between the samples and thus classification. Areas of the spectrum identified as important for classification included DNA regions, protein, amino acids, lipids and amide III and I. The PLSDA weights suggest that DNA peaks and amide III at 1257 cm^{-1} are more indicative of

lymphoma. The 934 cm^{-1} protein peak, phenylalanine, collagen (1312 cm^{-1}), amide III (1334 cm^{-1}), protein and phospholipid (1445 cm^{-1}), amino acids (1553, 1602 and 1613 cm^{-1}) and amide I (1648 cm^{-1}) are more indicative of *metastatic* measurements. The higher frequency of DNA peaks in the *lymphoma* measurements further indicate the presence of densely packed lymphocytes with stronger DNA signals. It would be of great interest to measure and analyse normal samples of both lymph nodes and the epithelial tissue from the organs where metastases often originate from, to gain a better understanding of whether the differences seen between the *lymphoma* and *metastatic* samples are as a result of different cancers or as a result of different tissue types. The peaks related to the biochemical differences between the cancers could then be identified more accurately.

When *lymphoma* samples were identified from only *non-cancer* samples, a sensitivity and specificity of 64% and 86% was calculated. Interestingly, the model that included all the data was slightly better at delineating *lymphoma* from *non-cancer* when grouped as *cancer* with the *metastatic* measurements. Peaks in the spectra that showed the most variation and were identified by VIP for delineating between the groups consisted of those recognised as more indicative of *lymphoma*, amino acid (656 cm^{-1}), cystine, protein (1408 cm^{-1}), protein and phospholipid peak around 1450 cm^{-1}, tryptophan (~1542 cm^{-1}), amino acid (1590 cm^{-1}) and amide I (1655 cm^{-1}). Important peaks more characteristic of *non-cancer* include, phenylalanine (623 and 1001 cm^{-1}), DNA (727, 746 and 781 cm^{-1}), carotenoid (1158 and 1521 cm^{-1}), tryptophan (1652 cm^{-1}) and amide I (1661 cm^{-1}). A higher level of DNA signal in the *non-cancer* spectra was not expected and should be more likely in the cancer samples as a result of increased cell production according to most studies. However, the mean spectra shows an elevated background in the averaged lymphoma spectrum when compared to *non-cancer* which might mask the DNA peaks also. A better understanding of the presence of the background would be needed to ascertain whether it is a useful feature for delineation or whether it would need further removal to reliably analyse pathological differences, but at the expense of less reproducible data and spectral integrity. Protein is recognised at stronger signals in the *lymphoma* data, indicating that perhaps there is a change in the nucleus to cytoplasm ratio. Also, there is a peaks in the PLSDA component weights that are both present in the

negative and positive direction. Another observation in the PLSDA weights was a higher signal of hydroxyapatite and cholesterol associated with calcifications. Calcifications are often found in malignant tissue but not as often in benign. Although it is reported that calcifications can also occur in the lymph nodes as a result of infection and may cause enlargement, leading to their removal. Higher carotenoid concentrations are in agreement with other studies and the findings in this research also.

Metastatic and *non-cancer* data were then analysed to assess the spectral differences between the two. A sensitivity and specificity were calculated as 75% and 97% for the identification of *metastatic* samples from *non-cancer*. The ability to distinguish *metastatic* from *non-cancer* was much better than for either *metastatic* and *lymphoma* or *lymphoma* and *non-cancer*. It is assumed that the higher level of variability between these two tissue types, is because of both different pathological classification as well as a difference in tissue type. *Lymphoma* and *metastatic* samples share the pathology of malignancy, though broad in its definition, and *lymphoma* and *non-cancer* share the same cell type. Important peaks for variation between the pathologies are those more attributed to *metastatic* measurements: tyrosine (851 cm^{-1}), phenylalanine (1001 and 1028 cm^{-1}), collagen (1310 cm^{-1}), the phospholipid and protein peak (~ 1445 cm^{-1}) and tryptophan (1551 cm^{-1}). Peaks in the spectrum more indicative of *non-cancer* measurements include: DNA (781, 1092, 1485 and 1574 cm^{-1}), carotenoid (1152 and 1517 cm^{-1}), tryptophan (1371 cm^{-1}) and amide I (1661 cm^{-1}). These results support the previous *metastatic* lymph node analysis where *metastatic* samples have less DNA and carotenoid signal and more protein and lipid than *non-cancer* samples. The weights in this part of the study showed collagen peaks to be indicative of both *non-cancer* and *metastatic* tissue measurements however. Such as at 883 and 1280 cm^{-1} in *non-cancer* LN tissue and 1270, 1294 and 1310 cm^{-1} in *metastatic* LN tissue. Nguyen *et al*.[276] show the difference between the Raman signals measured from type I and type IV collagen. There is a high degree of variability between the collagen types based on their structure. Different collagen types between the *non-cancer* and *metastatic* samples may lead to a variance in the signals measured and hence explain why collagen peaks are identified as being indicative of both *non-cancer* and *metastasis*. Alternatively, there could be shifts of the collagen peaks between the non-cancer and

metastatic measurements that contribute to variation, and thus lead to its presence in both the *non-cancer* and *metastatic* weights. In addition, Hilska *et al*.[277] observed a difference in the distribution of collagen types I, III, and IV, especially around glandular structures, between normal and cancerous tissue. Another explanation for the collagen presence in both *non-cancer* and *metastatic* tissue could be because of this change in distribution. This change may lead to a difference between the collagen types being detected based on where in the lymph node the measurement is taken from. However, only the collagen peak at 1310 cm^{-1} was included 100% of the time for classification.

Differences between classification sensitivities and specificities may result from a data imbalance between the two groups. A large difference between the sizes of datasets can lead to a preference of the model to classify a specific group as it is more represented. Thus, this may provide an explanation for the difference of 22% between the sensitivity and specificity of this cross-validation as there are more than four times the number of *non-cancer* samples as there are *metastatic*.

Analysis of the metastatic cancers showed no obvious spectral differences between the pathologies. PLSDA indicates there are more spectral differences between the *AC* and the *SCC* measurements than the *PC*. However, only one sample was measured from both *AC* and *PC* and that clearly limits the interpretation of this analysis. Whilst measuring the samples a shift in the phenylalanine peak was often observed at different pixels in the same map, which showed up in the weights. This could be as a result of the silicon offset used, where the phenylalanine band is detected between wavenumbers due to the limit of spectral resolution. Therefore, it may not be caused by an alteration in biochemical contributions although identified in the weights, as there is a limited number of samples in the group.

It would be of interest to spectrally compare the metastatic samples to their primary cancer sites to see if there are any similarities between them that could aid the diagnosis of unknown primary cancers (UPC). Gao and Li[278] found that the majority of research observed that morphological, cytological and molecular features of metastases with their matched primary site usually exhibit little differences. They noted that the theories behind the mechanism of metastasis

are often contradictory and much is still unknown. They also stated that research also found that metastases were more similar to the normal tissue of the organ they came from than the one they had settled in. They concluded that the inherent characteristics of malignancies are conservative, and whilst some differences occur between metastasis and site of origin these are minor and the majority of features remain similar. Therefore, their similarities might allow for primary site identification if more samples and a larger model could be produced. Knösel *et al*.[279] observed differences between lymph node and liver metastases, the former more likely to exhibit predominant tubular differentiation and a strong desmoplastic stroma reaction. These morphological changes they attribute to the biological behaviour and as a result of a higher presence of normal cells in the lymphatic system than the haematic.

There is a dramatic difference between the backgrounds of the *HL* and *NHL* spectra. Although the backgrounds of acquisitions alter slightly for every measurement they are normally not too different and vector normalisation tends to be able to reduce the difference even further. This change in background is most likely due to a difference in auto-fluorescence. This difference is more likely accentuated through measurements from only one sample in the *HL* group, rather than biochemical differences between the pathologies. More *HL* samples are needed to investigate further and make a more informed decision on the cause of this background change between the pathological groups.

Conclusion

This study shows the ability of Raman spectroscopy to recognise spectral differences between benign and malignant head and neck lymph nodes in a two group PLSDA model and identify primary and secondary cancer based on changes in their biochemical compositions. This shows the potential for Raman spectroscopy to be used as a diagnostic tool clinically to aid real-time in-situ analysis of lymph nodes, direct surgeries and reduce healthcare costs. The ability

to implement a better sensitivity using the ROC curve allows a lower number of false negatives, even if at the expense of more false positives.

The main findings include:

1. The main differences observed throughout the study between benign and malignant lymph nodes are a decrease in signal attributed to DNA, carotenoids, nucleic acids, some amino acids and an increase in lipids, proteins, collagen and some amino acids such as phenylalanine.
2. In the three group PLSDA model, lymphoma appears to be more similar to non-cancer measurements than to metastatic, indicated through its misclassification. This may suggest that benign and malignant lymphocyte populations are more similar than two separate malignant populations of different tissue types.
3. From the two group PLSDA model between lymphoma and metastatic, the confusion matrix shows that the metastatic group has a much lower sensitivity, suggesting the sample set is too small, metastatic tissue is less homogenous than lymphoma and often primary, or that the SCC, AC and PC are spectrally different leading to a poor ability to classify them as one individual group.
4. From analysis of the PLSDA and the weights, lymphoma appears to have more DNA, amide III and carotenoid assigned and less amino acid and lipid assigned peaks than metastatic.
5. Both the sensitivity and specificity increased by just over 10% when identifying metastatic samples from non-cancer, when compared with discriminating between lymphoma and non-cancer. This could be due to a higher level of biochemical similarities that leads to more similar spectral signals between the benign and malignant lymphocytes, regardless of pathology. Whilst phenylalanine is selected every time for classification, it is more indicative of non-cancer nodes when compared to lymphoma, yet less indicative when compared to metastatic.

A larger sample set of metastatic samples would be of great interest in taking this area of research further. The metastatic samples could also be subdivided, not only based on cell type, but also based on primary site. A much larger study would allow for more accurate inferences to be drawn and

for further hypotheses to be investigated. The collection of matched primary site tissue could also be useful to gain an understanding into any similarities to the metastatic site for diagnostic purposes for cases of UPC. Also, differences between the secondary and primary site would be interesting to analyse if any changes occur between them or whether biochemical changes are already present in certain cells that enable them to better metastasise and survive to set up a secondary site.

Chapter 6: Feasibility Study of a Raman Needle Probe

6.1 Results

Before lymph node measurements were performed using the needle probe, it was subjected to several provisional experiments to assess its performance and understand how best to measure biological samples. Intralipid and ink was used as a phantom to mimic the scattering and absorption properties of light in soft biological tissue, such as lymph nodes. India ink is often suspended in low absorbing diffusive media, such as lipid solutions like intralipid, in different quantities to obtain the desired absorption properties in a phantom to reflect that of actual samples. India ink acts as an absorber due to insoluble carbon particles suspended in the liquid. Its advantages for spectroscopic use is that it is spectroscopically stable, non-toxic, does not fluoresce and its absorption spectrum is smooth over the visible and NIR regions with little variation. However, it does not act only as an absorber but also as a scatterer making it difficult to obtain an accurate absorption coefficient.[280]

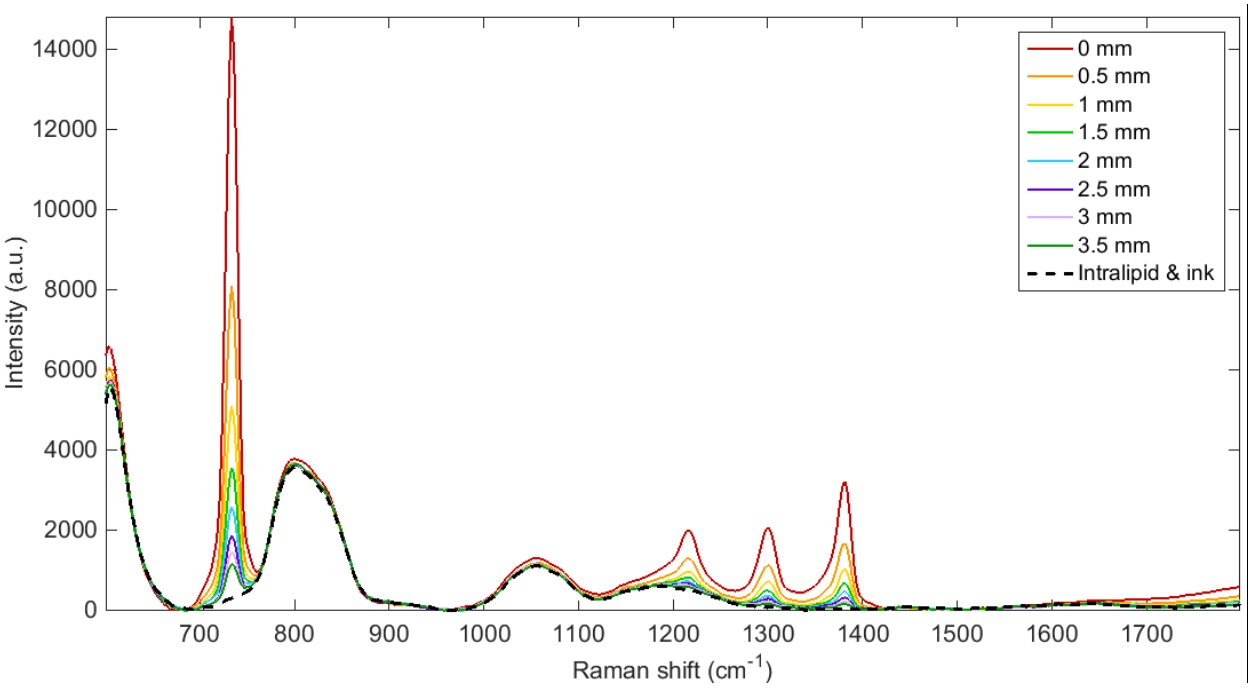

Figure 6.94- PTFE measurements in intralipid and ink solution measured at increasing distances of a 0.5 mm step size from the PTFE.

The PTFE measurements at different depths within the intralipid/ink phantom, displayed in Figure 6.94, shows the expected decrease in signal as the needle tip is pulled further away and measurements are taken at a greater distance from the sample. The strongest peak at 733 cm^{-1} is still observed until the last measurement was taken at 3.5 mm away. Additionally, the two peaks at 1299 and 1382 cm^{-1} are visible until 3.5 mm away, however the smaller peak at 1217 cm^{-1} is much more difficult to detect and is not as obvious.

For *in vivo* use, laser safety is an important consideration clinically. Although no burning was witnessed or measured when utilising the maximum output power of the laser (230 mW at the tip), spectral signals were acquired using just 20 mW of laser power. The time taken to take a measurement is also of surgical interest for faster diagnosis, both in the interest of treatment time and limiting patient discomfort.

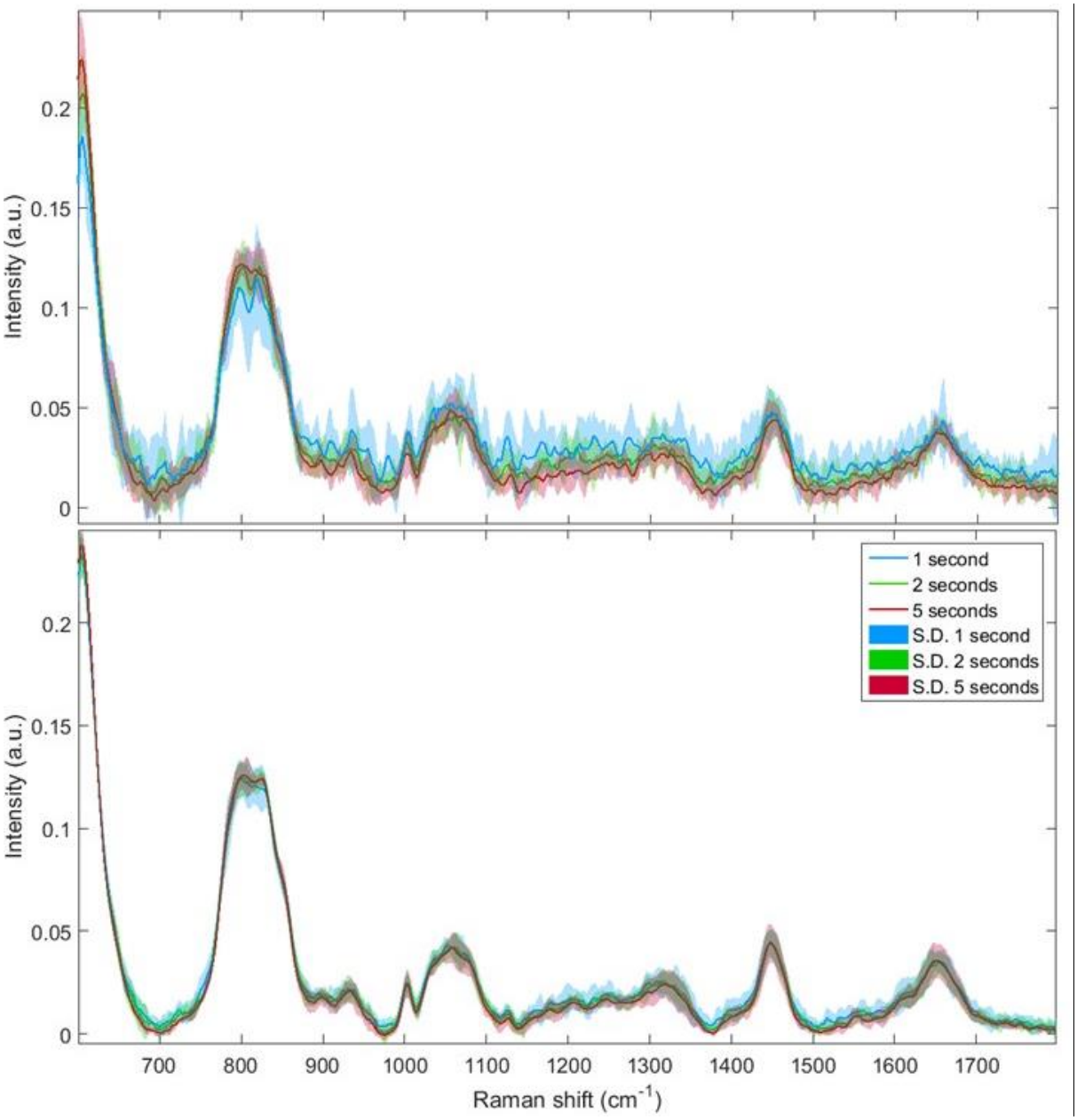

Figure 6.95- Measurement of pork meat with both 20 mW (top) and 230 mW (bottom) of power and measurement times of 1, 2 and 5 seconds with standard deviation plotted around the mean spectra.

Measurements of pork meat were taken using both 20 mW and 230 mW of laser power at 1, 2 and 5 seconds, displayed in Figure 6.95. Both sets of measurements at 20 mW and 230 mW at different times all have recognisable peaks throughout the spectra. Strong peaks at around 600 cm^{-1} and 800 cm^{-1} represent signal from the silica fibres of the system and are not biological. The system design uses optical filters to reduce this instrumental contribution, however the silica fibres still contribute to spectra and are responsible for these peaks. It is clear from the standard deviation that the measurements using 230

mW do not have as much variation between acquisitions as the 20 mW. The 1, 2 and 5 seconds spectra are also more similar to one another than the 20 mW measurements. The spectra are much smoother in the 230 mW spectra than the 20 mW. Lymph node measurements were eventually taken at 150 mW to get good signal whilst not probing with too much energy.

Figure 6.96 exhibits the spectra acquired from various tissues in two porcine samples. For visualisation, the data have been averaged across the ten measurements from each tissue type (5 from each sample) to represent the mean spectra for bone, fat, muscle and skin tissue in pork. Visual differences are present between the spectra. Stronger peaks representing the lipid contributions in the fat tissue at 1441 cm^{-1} and 1750 cm^{-1} and the peak around 960 cm^{-1}, assigned to the calcium-phosphate stretch[17] in bone tissue.

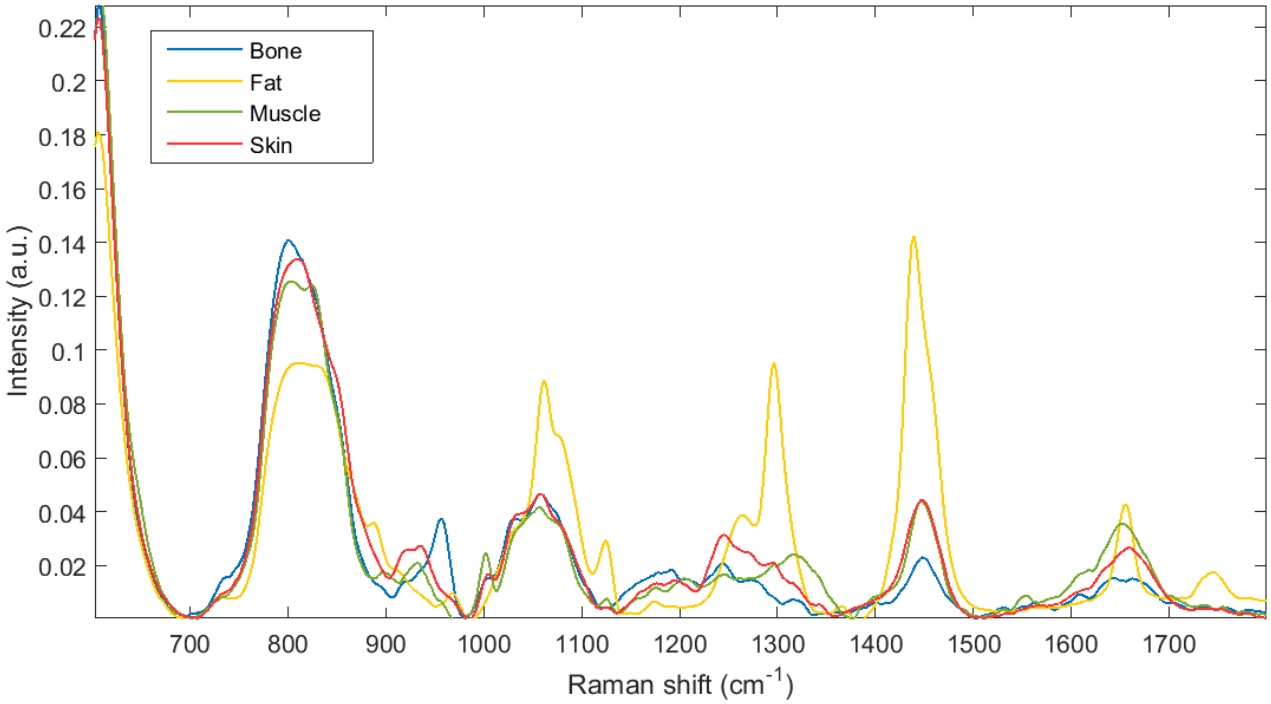

Figure 6.96- Mean spectra of bone (blue), fat (orange), muscle (green) and skin (red) averaged across 10 measurements each acquired over two pork samples.

Obvious grouping and separation is seen in the PLSDA score plot, Figure 6.97, across components 1, 2 and 3. This was further supported through a 10-fold cross-validation which yielded a preliminary classification accuracy of 100%,

where four spectra, one from each group, were omitted from the training model to blind test and predict classifications.

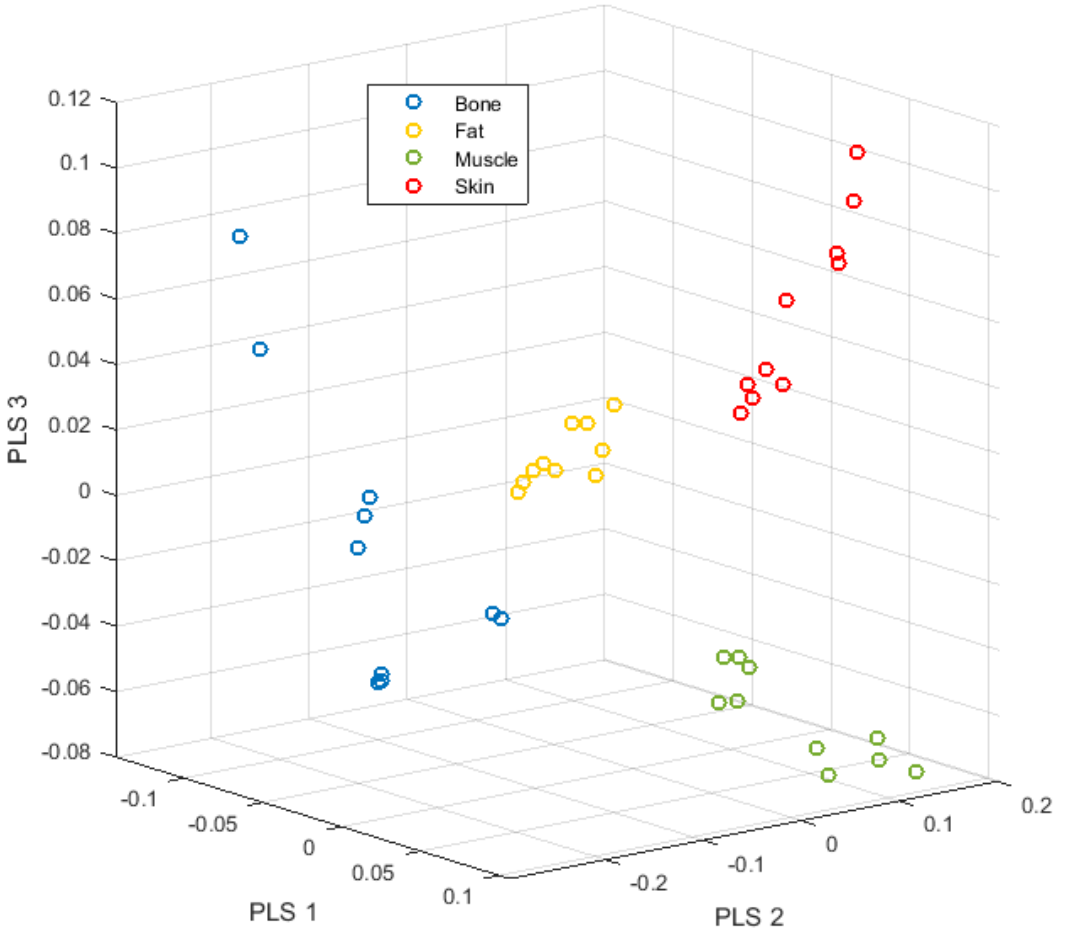

Figure 6.97- PLSDA scores plot for bone (blue), fat (yellow), muscle (green) and skin (red) measurements.

The weights in Figure 6.98, show the peaks in the spectra that contribute the most variance between the groups. The biochemical peaks are primarily responsible for the separation between groups rather than any spectral noise, signal or baseline. This shows that the needle probe is sensitive enough to measure signal attributed to different biochemical signals within tissue samples, to identify them as variable in the PLSDA model. Also, coupled with MVA, unseen spectra can be blind tested and correctly classified, in this small initial feasibility study for testing the needle probes validity for pathological samples.

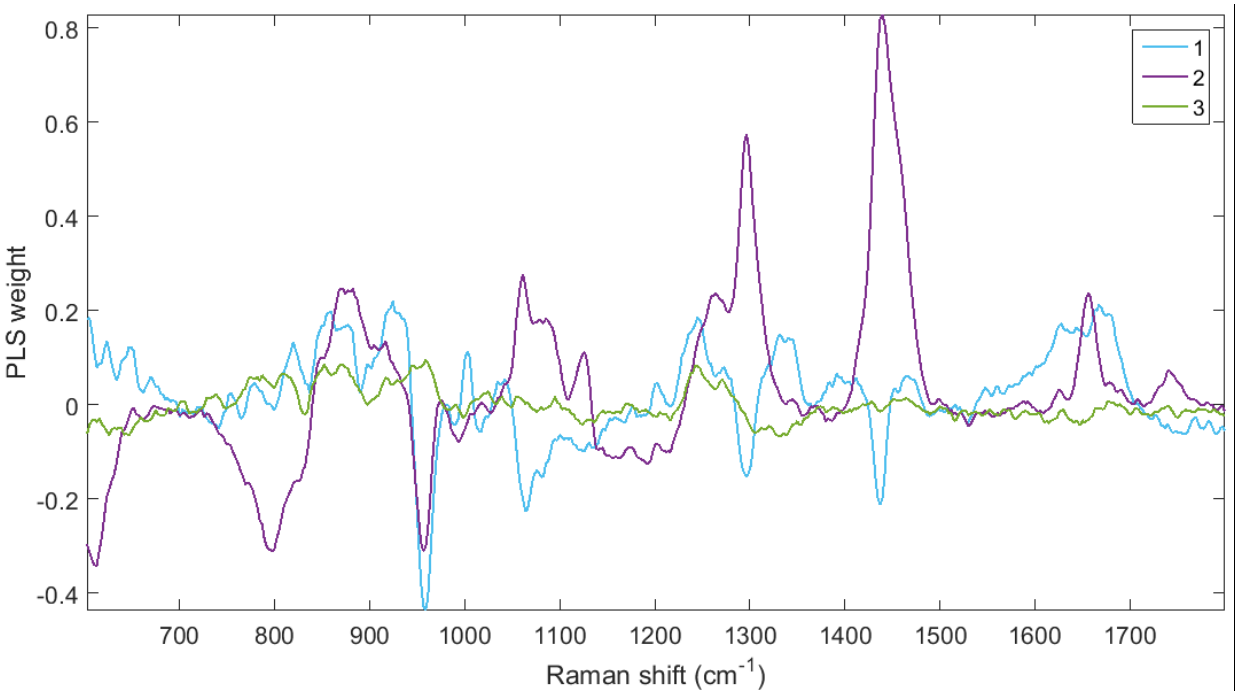

Figure 6.98- PSDA weight from the top three components for the model of pork bone, fat, muscle and skin.

To assess the ability of the Raman needle system to distinguish between more subtle biochemical differences, the skeletal muscle across four different animal species were measured at an acquisition time of 1 second and analysed using MVA. The muscle spectra from, beef, chicken, lamb and pork meat were obtained. Figure 6.99 shows data averaged over 10 spectra for ease of visualisation. The differences between the spectra from the different tissue samples are more subtle than between the spectra from different tissue types within pork meat. The gross quantities of many proteins, lipids and nucleic acids are similar throughout the different samples of skeletal muscles and are reflected in similar peak intensities and ratios in their spectra. Similar biochemical concentrations within the tissue samples are most likely due to the functions required of the muscles being similar across the animals measured.

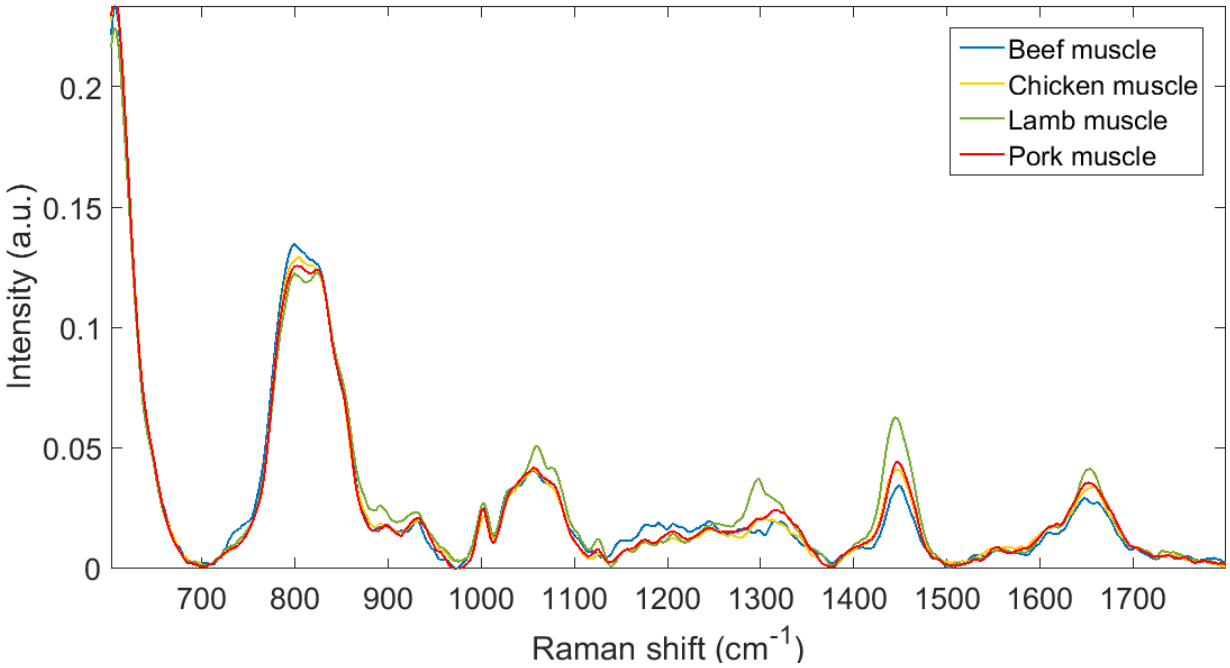

Figure 6.99- Mean spectra averaged across 10 muscle measurements of beef (blue), chicken (orange), lamb (green) and pork (red) samples.

The PLSDA scores plot, Figure 6.100, shows the grouping of the different muscle tissues and separation between the different species. The measurements are clearly grouped but their separation is less obvious with some slight overlap. As a result, the classification accuracy, calculated from 10-fold cross-validation, reflects this lack of complete delineation and was calculated to be 97.5% on average. The weights of the PLS components, Figure 6.101, show the peaks in the spectrum which are considered to contribute the most to the variance between the groups. The first and second components contain most of the spectral variance attributed to biochemicals whilst the third component is heavily composed of peaks from spectral noise.

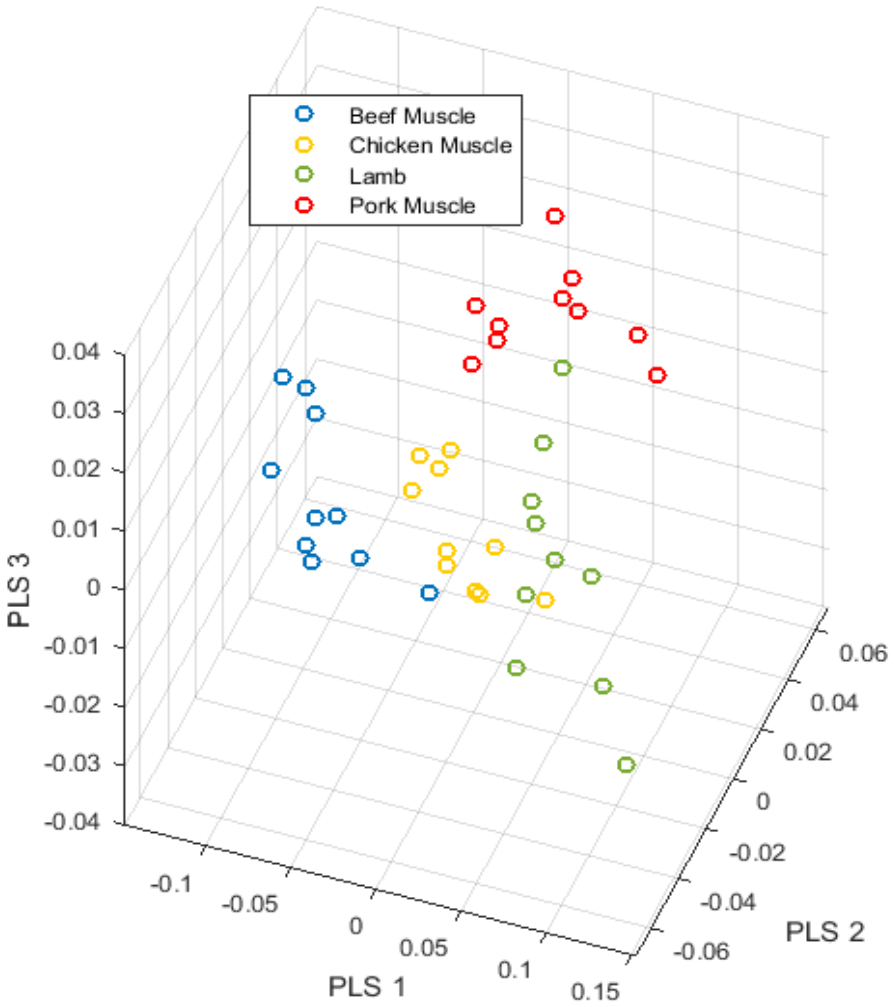

Figure 6.100- PLSDA scores plot of muscle measurements of beef (blue), chicken (yellow), lamb (green) and pork (red).

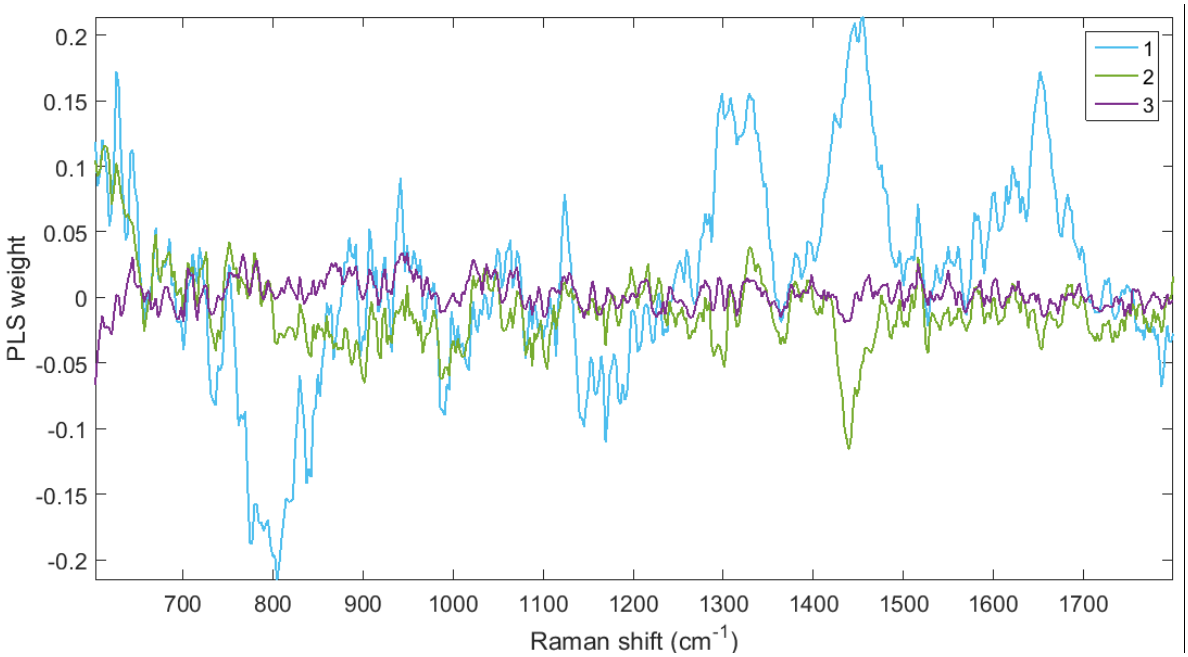

Figure 6.101- PLSDA weights from the first three components of the beef, chicken, lamb and pork model.

The lymph node samples obtained for this study were much smaller than the bulk animal tissues samples, usually measuring less than 10 mm in diameter. To examine the suitability of the signal gained using the needle probe, one bulk non-cancer oesophageal lymph node was analysed after the section had been obtained for inVia measurements. Both the lymph node tissue and the surrounding adipose tissue was measured to provisionally ensure that the spectral quality was good enough to observe differences between these two different tissue types so the correct tissue was measured when the lymph nodes are harder to detect amongst the adipose. The biochemical differences between lymph node tissue and adipose tissue is evident, observed in their spectra in Figure 6.102. Peaks attributed to lipids, at 1068, 1123, 1303, 1441 and 1750 cm^{-1}, are more intense. Those attributed to amino acids and DNA, such as the phenylalanine peak at 1002 cm^{-1} and the DNA peak at 824 cm^{-1}, are weaker in the adipose tissue than the lymph node tissue. MVA was not carried out on the oesophageal lymph node data as only one sample was investigated and a few spectra taken due to limited sampling areas due to the nodes small size.

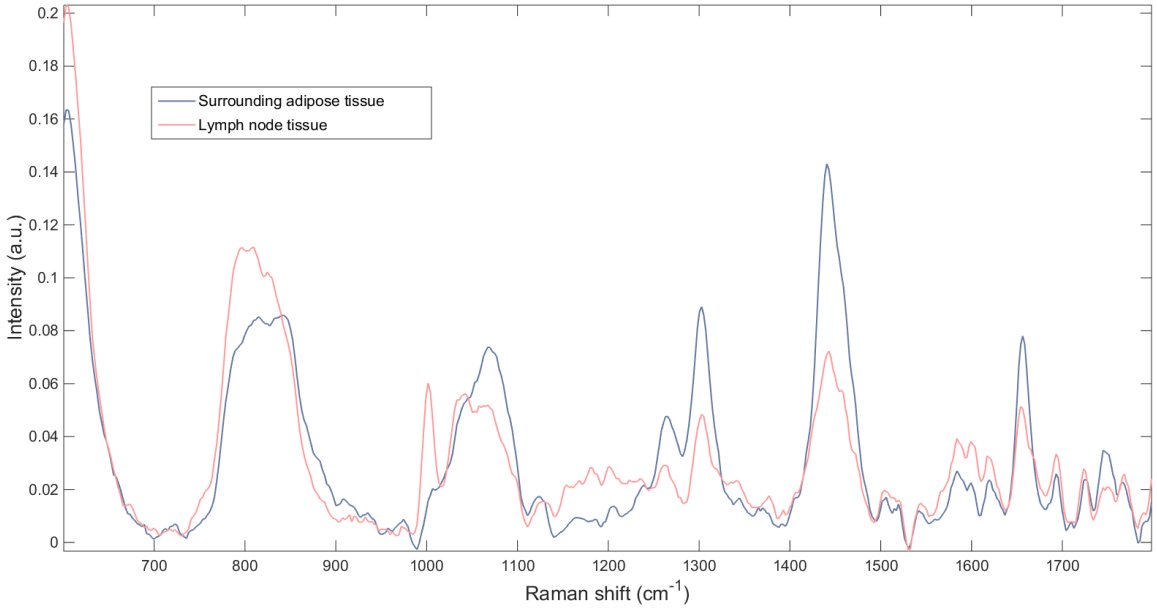

Figure 6.102- Mean spectra of lymph node tissue and the surrounding adipose averaged over 2 measurements.

After provisionally assessing the Raman needle probes proficiency to measure lymph node samples, the bulk head and neck lymph nodes were then analysed. The samples were usually half of the lymph node sample excised. Occasionally in the metastatic cases where many lymph nodes were taken, a whole node might be used. Therefore, the samples were not as large as a whole malignant lymph node, again limiting the sampling area. Measurements were taken from areas over the whole sample to collect representative data. The mean spectra of the cancer and non-cancer measurements, Figure 6.103, show visual spectral differences between the two groups.

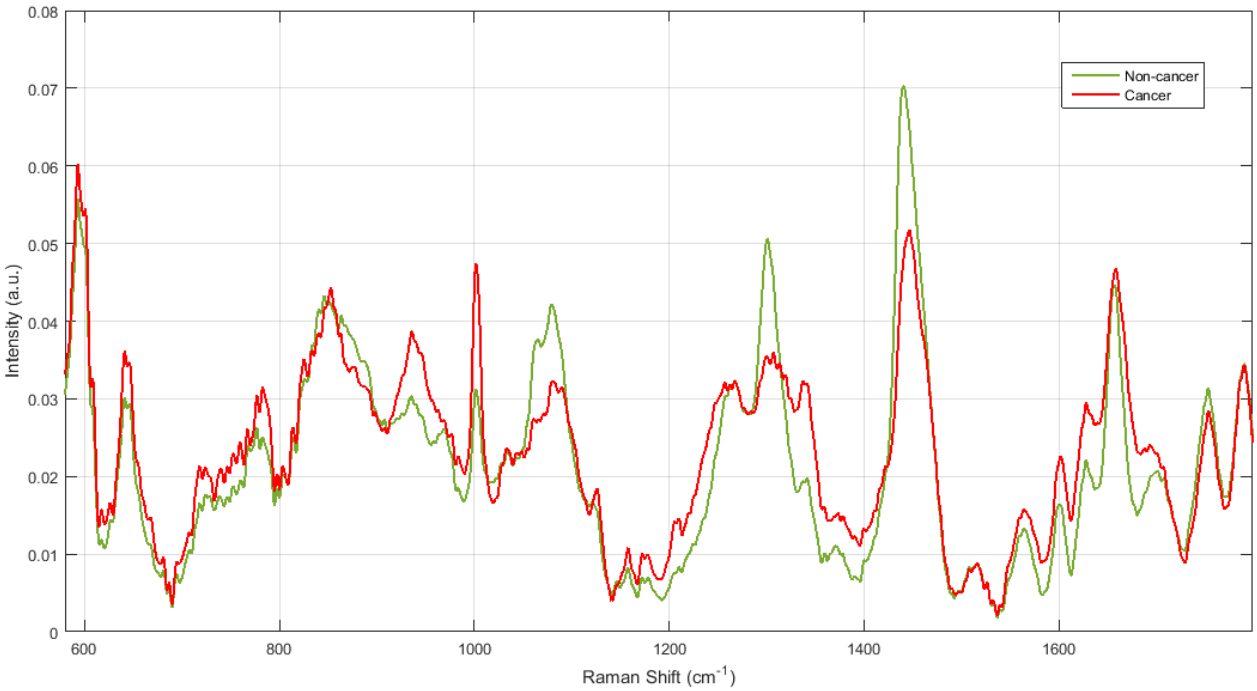

Figure 6.103- Mean spectra of the non-cancer (green) and cancer (red) lymph node measurements.

PLSDA was performed on the data and the scores plot can be seen in Figure 6.104. Separation of the data based on their groups is present primarily in component 1. Neither component 2 or 3 appears to separate the groups visually. The non-cancer measurements are more dispersed across these components whilst the cancer data appear closer to 0. There are more non-cancer data and samples and therefore dispersion is a more likely occurrence due to a higher level of natural variation. Cross-validation calculated the sensitivity and specificity of identifying cancer lymph node measurements from non-cancer as 79% and 81% (Table 6.50).

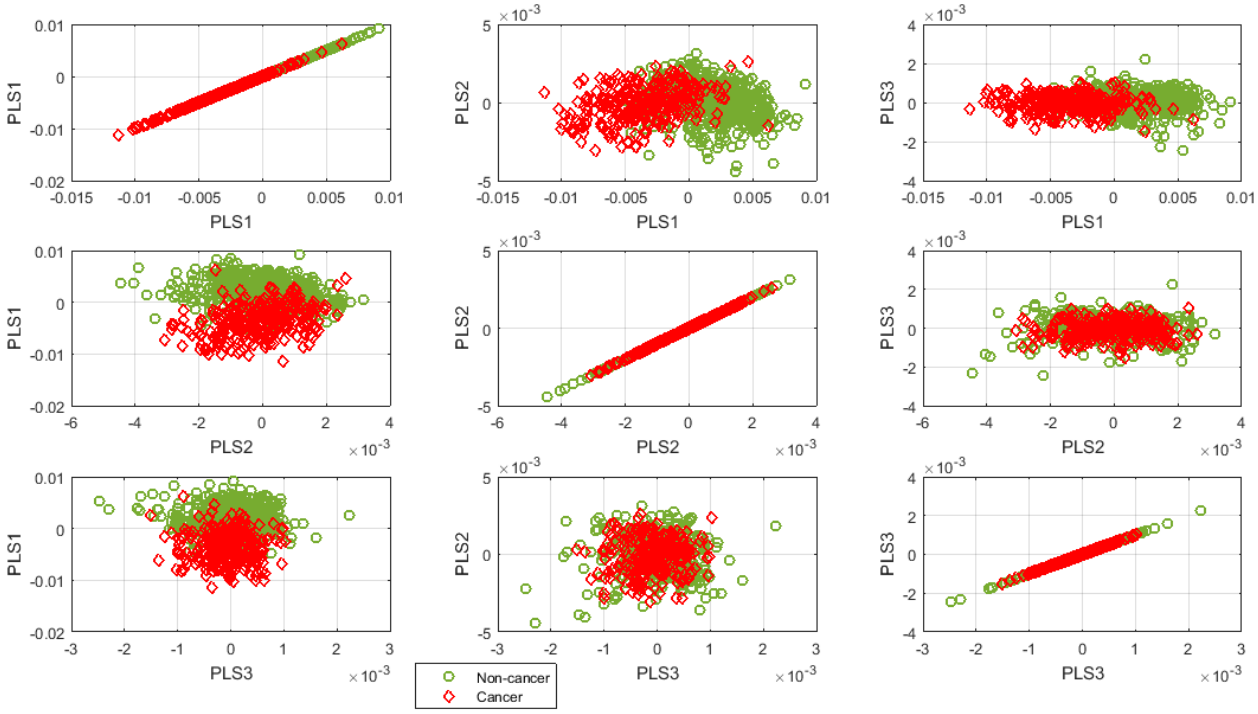

Figure 6.104- PLSDA scores plots of the top three components of the cancer and non-cancer measurements.

Table 6.50- Confusion matrix of from the cross-validation of the PLSDA model of cancer and non-cancer.

		True group	
		Non-Cancer	**Cancer**
Predicted group	**Non-Cancer**	29	3
	Cancer	7	11
	Total Samples	36	14
	Sensitivity	0.8056	0.7857
	Specificity	0.7857	0.8056

The ROC curve, Figure 6.105, shows that to achieve a 90% and 95% sensitivity, a compromised specificity of 36% and 23% would occur. However clinically, a higher sensitivity may be of higher importance to avoid false negatives whilst still reducing the number of unnecessary biopsies.

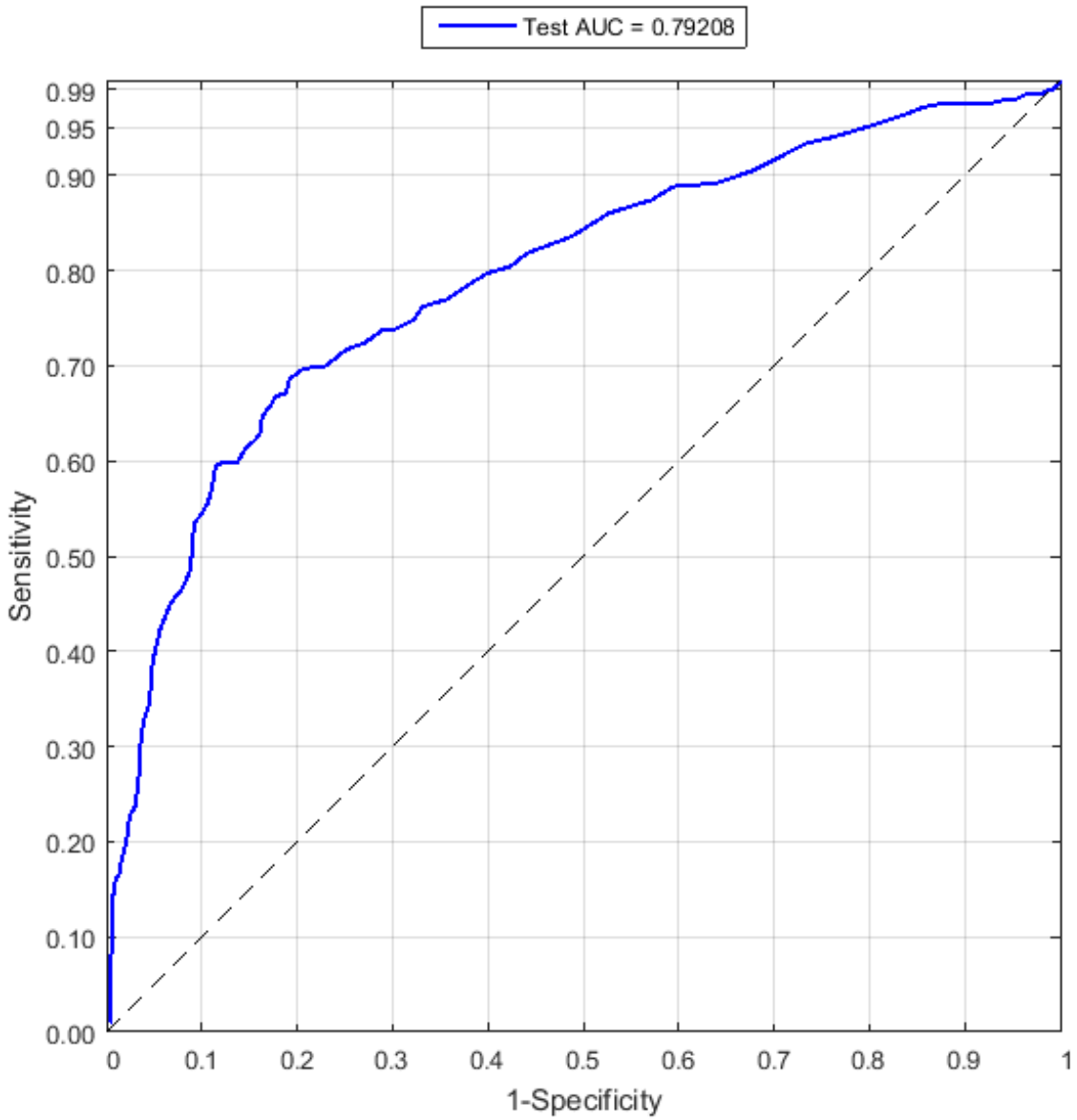

Figure 6.105- ROC curve for the classification of non-cancer and cancer measurements.

Figure 6.106 shows the etaloning of the CCD in the green glass measurements. The green glass, though measured on different days, can be seen to exhibit a very similar pattern, thus this effect is relatively consistent. However, the etaloning is not consistent enough to simply correct for it during pre-processing and much more work is needed to reproducibly and correctly eliminate its effect from the spectra, and not interfere with the biochemical signal.

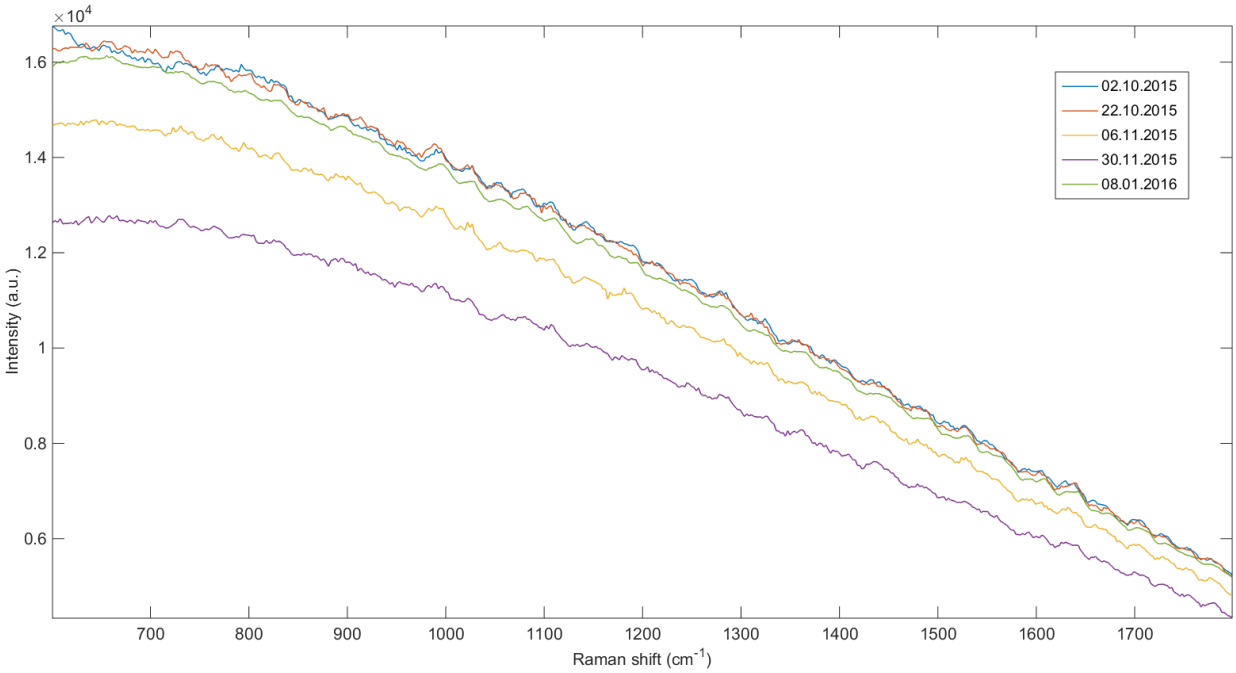

Figure 6.106- Green glass measurements across different days showing the effect of etaloning.

6.2 Discussion

The phantom measurements were able to provide a preliminary indication of the sampling volume that could be measured using the Raman needle probe. A distance of 3.5 mm from the tip still suitable to detect the peaks of PTFE. Normal lymph nodes are usually up to 1 cm in diameter, therefore a large proportion of the node could be sampled using this needle probe. However, PTFE has stronger signals than biological samples, so while the sampling area might be large, the signal might be much weaker. Matousek *et al*.[151] showed that the spectral signature of bone can be detected at 3 mm distance using spatially offset Raman spectroscopy (SORS) but believed it could be increased to 5 mm and more, limited only by the lack of inelastic scattered photons detected.

Although Raman spectroscopy utilises non-ionising radiation and considered safer than many current diagnostic techniques, such as CT, it can still result in thermal burning. There is currently a lack of consensus surrounding an acceptable power limit. The difference in wavelengths of light used and area probed, contributes more variables that further complicates a solution for the standard safe exposure limit. The tissue type irradiated also contributes to the

difference in response observed. Bergholt *et al*.[281] found that using a 785 nm laser at 30 mW with a 200 µm spot size could only increase the temperature of a sample by a maximum of 0.15 ^{o}C when probing a tissue sample for at least a minute. They stated that the temperature rise is far below that which could cause cell cytotoxicity. This study showed that both a laser power of 60 mW and 230 mW at the sample produced biological spectra with identifiable peaks at 1, 2 and 5 second acquisitions. However, the main difference between the two laser powers was the reduced noise when using 230 mW. A laser power of 150 mW was used for the rest of the study to provide a mid-range power that was both enough to reduce noise contributions but also probe the tissue without causing concern regarding tissue burning. A more in-depth study regarding the laser power and acquisition time would be useful in order to optimise measurement quality, measurement time and patient safety.

Another issue that has been observed is the influence of different light sources on measurements and their resulting contamination of the spectra. As Raman scatering is a weak phenomenon, the resulting spectra of biological samples can easily be contaminated by the detection of ambient light and is thus better carried out in the dark to reduce or remove this effect. In theatre settings, lights are essential in order to carry out surgery.[282] Nyberg *et al.*[282] noted the presence of three ambient fluorescent light peaks 985 cm^{-1}–993 cm^{-1}, 1343 cm^{-1}–1353 cm^{-1}, 1537 cm^{-1} and 1547 cm^{-1}. However, they detected and removed these peaks by replacing the values with those ot the neighbouring pixels. Nyberg *et al.*[283] combatted the issue, through probe insertion of at least 1 mm into tissue samples and acquiring spectra in the dark, to reduce ambient light contribtutions in measured spectra.

Desroches *et al.*[284] studied the effect of common operating room (OR) light sources on measured spectra. Through measurements taken of veal and the inclusion of various light powers and sources both indirectly and directly, coupled with principal component analysis, they found that whilst some light sources could be kept on during measurements, others had too much of an effect on the spectra. They concluded that indirect OR lights, liquid crystal display (LCD) screens and exterior radiation entering from windows do not need to be eliminated whilst taking measurements. However, microscope, fluorescent and

direct OR lights cannot be used without significantly affecting the resulting spectra. They also observed that whilst IR sources do not contribute distinguishable peaks to measured spectra and any contribution can be reduced through background correction algorithms, they can result in detector saturation. Therefore, IR sources must be pointed away from the spectroscopic probe during measurements to avoid this problem. They note the importance and encourage specific OR measurements to be carried out prior to using probe spectrometers in surgery as different light sources and intensities may vary the spectral contributions experienced whilst taking measurements. They stress the importance of situating PC monitors and other LCD screens away from the area where measurements take place to avoid spectral artefacts that could occur.

The preliminary biological tissue study carried out on different animal samples demonstrated the feasibility of the Raman needle probe for its use on human tissue. The ability of the probe to detect the spectral differences between the different tissue types in two porcine samples is demonstrated both through visual variations in the spectra but also after multivariate analysis. The probe is also capable of detecting more subtle biochemical changes between muscle samples from the different animals, less visually obvious in the spectra, but shown via multivariate analysis.

The initial set of lymph node measurements have shown a promising use of the needle probe as a surgical tool, through very clear spectral differences between adipose tissue and lymph node tissue based on the biochemical makeup of the two tissue types. Lymph nodes are typically firmly situated in adipose tissue.[231] The lymphatic fluid contains large amounts of emulsified lipids, and fat deposition often occurs initially along lymphatic structures. Lymphatic fluid leakage from lymphatic vessels appears to promote fat accumulation.[232] An increased number of adipocytes surrounding the lymph nodes is frequently associated with chronic inflammation of nodes.[233] Lymph nodes resected during surgery are often removed due to their enlarged appearance and therefore may be surrounded by a large concentration of adipocytes.[242] Surgeons often mistake adipose tissue for fat when taking a biopsy. Whilst collecting head and neck lymph node samples for this study, 5 out of the 84 collected samples turned out to be tissue other than lymph node when analysed by a pathologist, three were

fat and the other two were connective tissue. The obvious spectral differences observed between these two tissue types could allow real-time surgical guidance to help locate lymph nodes during surgery and avoid the unnecessary removal of fat tissue and wasted time in pathology.

The measurement and analysis of the head and neck lymph node samples resulted in a 79% and 81% sensitivity and specificity for the classification of cancer from non-cancer lymph node samples using a leave-one-sample-out cross-validation. This is similar to the inVia Raman measurements of 77% and 89% and shows its comparability to *ex vivo* techniques. This feasibility study shows that the needle probe can measure biological tissue samples at a good sensitivity, able to differentiate between pathological types using MVA based on subtle spectral differences.

Although the needle probe has a retracting needle tip to allow minimum debris build-up, it will still be inserted through skin, muscles and adipose tissue before reaching the lymph node. Thus, the collection of tissue debris should still be examined through extensive experimentation where the needle probe is inserted through these different tissues to assess whether this could be a problem or whether the retracting tip provides enough protection.

As a result of etaloning, and the inability to fully correct for its contributions and those of the instrument, interpretation of the spectra are limited and biochemical inferences were not made at this stage. Etaloning is caused by the interaction of the light with the front and back silicon surfaces of the back thinned deep depletion CCD, resulting in its reflectance back and forth. The reflected light interferes with itself and an interference pattern in the spectrum is then produced. A different laser wavelength or CCD chip would have an altered etalon pattern and produce spectra that reflect this. As a result, in order to use multiple units, the etaloning needs to be removed from spectra, either through the use of a different CCD that does not produce this phenomenon or through pre-processing, in order to correctly interpret measurements and assess its clinical use. Andor[285] suggest alternative detectors that limit this, such as those that have deeper depletion regions to avoid re-reflectance of the light. Front illuminated CCDs are also suggested as an alternative, although these have lower sensitivity due to

photon interaction with the polysilicon electron structure. A radiometric calibration may be applied, binning to potentially average out the etalon signal or a fringe-suppression to the CCD where the back wall is roughened to avoid re-reflectance may also help to reduce this phenomenon in spectra.

6.3 Conclusions

This feasibility study has shown the ability of this Raman needle probe for measuring biological samples and has demonstrated its comparability to current techniques for the blind identification of pathology using MVA. The potential for *in vivo* use and as a clinical tool has also been shown in this study. The speed at which measurements can be taken and the quality of spectral acquisitions demonstrate the usefulness of this technique both clinically and scientifically.

Further work should be carried out to correct for etaloning and instrument contributions to better interpret measurements. A larger study would also be useful to gain a better indication of the diagnostic potential. The ability to differentiate between the different types of cancer within the lymph node would also be of interest to compare the technique to current *ex vivo* Raman methods and assess if there are any limitations to the use of the probe. A clinical trial would be needed to implement the technique in clinics and laboratories to offer a tool to guide surgeries and initially assess samples in pathology laboratories.

Conclusions and Future Work

This study has successfully demonstrated and evaluated the ability of Raman spectroscopy for the detection of biochemical changes between cancer and benign tissue from lymph nodes of different areas.

A sensitivity and specificity of 71% and 89% was calculated for identifying metastatic from non-cancer lymph nodes for oesophageal, femoral and head and neck lymph nodes collectively. Primary differences between malignancy and non-cancer appear to be due to lower DNA and carotenoid and higher lipid and collagen concentrations indicated through variations in their assigned peak intensities. PLSDA and cross-validation of the separate lymph node regions led to sensitivities of 50%, 83% and 80% and specificities of 83%, 75% and 92% for identification of oesophageal, femoral and head and neck metastases from non-cancer lymph nodes respectively. A much lower sensitivity in the oesophageal lymph nodes analysis may have been influenced by a large level of carbon particle deposits and strong lipid peak contributions, resulting difficulty measuring and analysing samples. Whilst femoral and head and neck nodes can be better classified individually, oesophageal classification may be improved when compared against data from the other lymph node regions. The spectral maps taken from the same head and neck nodes were analysed and produced the same sensitivity and specificity as the PLSDA model of point acquisitions. This provides support for the ability of classifying *in vivo* measurements based on pathology, where sampling would be limited to a few point acquisitions.

The larger head and neck lymph node study, analysing the classification capabilities of both metastatic and primary lymphomas, resulted in sensitivities of 64% and 75% and specificities of 86% and 97% for identifying lymphoma and metastatic samples respectively from non-cancer samples. The sensitivity and specificity of classification for cancer overall was 82% and 90%. The needle probe measurements demonstrated the feasibility for *in vivo* use. The sensitivity and specificity for classification of the needle probe data using the bulk lymph node of the same samples as the sections were 81% and 79%. A similar accuracy to the confocal Raman section measurements shows the functionality of the

needle probe for analysing subtle biochemical changes and indicates its potential as a clinical tool for real-time diagnosis and as a surgical assistant.

This study has resulted in the development of further hypotheses that should be investigated to benefit this area of research. The main limitation of this study is the sample size, and therefore conclusions cannot be drawn based on being representative of the general population.

The collaboration of different research centres and hospitals would enable the collection of a larger sample-set, likely to represent a higher number of pathologies, more measurements and a faster output of data. Larger studies are more able to pave the way for clinical trial. Therefore, after the issues with the needle probe are overcome and a larger *ex vivo* trial is found to be successful for accurate classification of lymph nodes, its use clinically, *in vivo*, would be an obvious next step. A clinical trial would introduce the technique into hospitals and enable analysis of the diagnostic potential and its use as a surgical tool.

The presence of benign diseases is a frequent occurrence with lymph node biopsies. Hence, a study utilising Raman spectroscopy for the identification of different diseases would be advantageous to support the diagnostic potentials of the technique for different lymph node pathologies. When a patient is first seen by a physician regarding a suspected cancer, it is usually as a result of a noticeably inflamed lymph node or as a result of suspected metastasis from a primary cancer. In both cases the cancer would be at a further stage of progression than a pre-cancer. Therefore a Raman needle probe is more likely to investigate these cancers, where they would be more discernible from a non-cancerous lymph node. It would be of great interest to investigate pre-cancers and early-stage cancers of the lymph node to determine the suitability of a needle probe as a screening method for primary cancers. This would enable better patient care from clinicians and allow the early removal of any tissue that may pose a metastatic threat.

Overall this study provides a thorough investigation into the use of Raman spectroscopy for the analysis of metastatic and primary malignancies found in various the lymph nodes from different anatomical locations. Furthermore it

demonstrates the potential for the *in vivo* use of Raman spectroscopy for diagnostic and surgical purposes. The technique contributes to the fast-growing field through expanding on research previously undertaken and providing the introduction and analysis of a novel technique that delivers the technology needed to support clinical implementation.

References

1 Cancer Research UK, .

2 Cancer research UK, .

3 D. I. Ellis and R. Goodacre, *Analyst*, 2006, **131**, 875–85.

4 C. Kendall, M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker and N. Stone, *Analyst*, 2009, **134**, 1029–45.

5 G. K. Hahm, T. H. Niemann, J. G. Lucas and W. L. Frankel, *Arch. Pathol. Lab. Med.*, 2001, **125**, 736–9.

6 N. Stone, C. Kendall, N. Shepherd, P. Crow and H. Barr, *J. Raman Spectrosc.*, 2002, **33**, 564–573.

7 J. Moros, S. Garrigues and M. D. La Guardia, *TrAC Trends Anal. Chem.*, 2010, **29**, 578–591.

8 L.-P. Choo-Smith, H. G. M. Edwards, H. P. Endtz, J. M. Kros, F. Heule, H. Barr, J. S. Robinson, H. a Bruining and G. J. Puppels, *Biopolymers*, 2002, **67**, 1–9.

9 B. Brożek-Płuska, I. Placek, K. Kurczewski, Z. Morawiec, M. Tazbir and H. Abramczyk, *J. Mol. Liq.*, 2008, **141**, 145–148.

10 Cancer Research UK, 2010.

11 M. Watson, A. Berrett, R. Spence and C. Twelves, *Oncology*, Oxford University Press, 2nd edn., 2006.

12 L. Pecorino, *Molecular Biology of Cancer*, Oxford University Press, 3rd edn., 2012.

13 I. ur Rehman, Z. Movasaghi and S. Rehman, *Vibrational Spectroscopy for Tissue Analysis*, Taylor and Francis Group LLC, Florida, 2013.

14 Cancer Research UK, 2014.

15 I. F. Tannock, R. P. Hill, R. G. Bristow and L. Harrington, *The Basic Science of Oncology*, The McGraw-Hill Companies, Inc., 4th edn., 2005.

16 M. Andreeff, D. W. Goodrich and A. B. Pardee, in *Holland-Frei Cancer Medicine*, eds. R. C. J. Bast, D. W. Kufe and R. E. Pollock, Hamilton (ON): BC Decker, 5th edn., 2000.

17 L. E. Donate and M. A. Blasco, *Philos. Trans. R. Soc. B Biol. Sci.*, 2011, **366**, 76–84.

18 D. Hanahan and R. a Weinberg, *Cell*, 2011, **144**, 646–74.

19 R. J. B. King and M. W. Robins, *Cancer Biology*, Pearson Education Limited., 3rd edn., 2006.
20 T. L. Vincent and R. a Gatenby, *Int. J. Oncol.*, 2008, **32**, 729–37.
21 A. Weston, C. Curtis and M. D. Harris, in *Holland-Frei Cancer Medicine.*, ed. D. W. K. and R. E. P. Bast, R.C. Jr., Hamilton (ON): BC Decker, 2000.
22 Y. J. Moon, X. Wang and M. E. Morris, *Toxicol. In Vitro*, 2006, **20**, 187–210.
23 K. S. Polyzoidis, G. Miliaras and N. Pavlidis, *Cancer Treat. Rev.*, 2005, **31**, 247–255.
24 M. S. Pepper, *Clin. Cancer Res.*, 2001, **7**, 462–468.
25 S. Wong and R. Hynes, *Cell Cycle*, 2006, **5**, 812–817.
26 I. J. Fidler, *Nat. Rev. Cancer*, 2003, **3**, 453–458.
27 P. Nowell, *Science (80-.).*, 1976, **194**, 23–28.
28 W. C. Mueller, Y. Spector, T. B. Edmonston, B. St Cyr, D. Jaeger, U. Lass, R. Aharonov, S. Rosenwald and A. Chajut, *Oncologist*, 2011, **16**, 165–74.
29 D. Tarin, *Tissue interaction in carcinogenesis*, Academic Press Inc. LTD., London, 1972.
30 A. Mahadevan-Jansen and R. R. Richards-Kortum, *J. Biomed. Opt.*, 1996, **1**, 31–70.
31 M. Bower and J. Waxman, *Lecture Notes: Oncology*, Wiley-Blackwell, 2nd edn., 2011.
32 C. L. Willard-Mack, *Toxicol. Pathol.*, 2006, **34**, 409–24.
33 N. L. Harris, H. Stein, S. E. Coupland, M. Hummel, R. D. Favera, L. Pasqualucci and W. C. Chan, *Hematology Am. Soc. Hematol. Educ. Program*, 2001, **2001**, 194–220.
34 Cancer Research UK, 2013.
35 N. A. Campbell, J. B. Reece, L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky and R. B. Jackson, *Biology*, Perason Education Inc., 8th edn., 2008.
36 C. J. McMahon, N. M. Rofsky and I. Pedrosa, *Radiology*, 2010, **254**, 31–46.
37 D. L. Drayton, S. Liao, R. H. Mounzer and N. H. Ruddle, *Nat. Immunol.*, 2006, **7**, 344–353.
38 C. Hadamitzky, H. Spohr, A. S. Debertin, S. Guddat, M. Tsokos and R. Pabst, *J. Anat.*, 2010, **216**, 556–62.
39 T. J. Kipps, in *Williams Hematology*, eds. K. Kaushanky, M. A. Lichtman,

E. Beutler, T. J. Kipps, U. Seligsohn and J. T. Prchal, The McGraw-Hill Companies, Inc., 8th edn., 2012.

40 B. D. Cheson, S. J. Horning, B. Coiffer, M. A. Shipp, R. I. Fisher, J. M. Connors, T. A. Lister, J. Vose, A. Grillo-Lopez, A. Hagenbeek, F. Cabanillas, D. Kippensten, W. Hiddermann, R. Castellino, N. L. Harris, J. O. Armitage, W. Carter, R. Hoppe and G. P. Canellos, *J. Clin. Oncol.*, 1999, **17**, 1244–1253.

41 D. Mosser and J. Edwards, *Nat. Rev. Immunol.*, 2008, **8**, 958–969.

42 P. J. Murray and T. a Wynn, *Nat. Rev. Immunol.*, 2011, **11**, 723–37.

43 E. Steliarova-Foucher, C. Stiller, B. Lacour and P. Kaatsch, *Cancer*, 2005, **103**, 1457–67.

44 A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt, *Nature*, 2000, **403**, 503–511.

45 Cancer Research UK, 2015.

46 S. Kemp, G. Gallagher, S. Kabani, V. Noonan and C. O'Hara, *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.*, 2008, **105**, 194–201.

47 S. Yamamoto, H. Nakase, K. Yamashita, M. Matsuura, M. Takada, C. Kawanami and T. Chiba, *J. Gastroenterol.*, 2012, **45**, 370–388.

48 E. Zucca and F. Cavalli, *Ann. Oncol.*, 2000, **11**, 219–22.

49 N. L. Harris, E. S. Jaffe, G. Diebold, J.Flandrin, H. K. Muller-Hermelink, J. Vardiman, T. A. Lister and C. D. Bloomfield, *Am. J. Clin. Pathol.*, 1999, **17**, 3835–3849.

50 F. L. Greene, D. L. Page, I. D. Fleming, A. G. Fritz, C. M. Balch, D. G. Haller and M. Morrow, *AJCC Cancer Staging Manual*, American Joint Committee on Cancer, 6th edn., 2001.

51 P. G. Isaacson, *Br. J. Haematol.*, 2000, **109**, 258–266.

52 C. Berard and R. Hutchison, *Ann. Oncol.*, 1997, **8**, S3–S9.

53 M. Pfreundschuh, *Blood*, 2010, **116**, 5103–10.

54 H. Mani and E. Jaffe, *Clin. Lymphoma Myeloma*, 2009, **9**, 206–216.

55 T. C. Kwee, R. M. Kwee and R. a J. Nievelstein, *Blood*, 2008, **111**, 504–16.

56 P. P. Carbone, H. S. Kaplan, K. Musshoff, D. W. Smithers and M. Tubiana, *Cancer Res.*, 1971, **31**, 1860–1861.

57 S. a Stacker, S. P. Williams, T. Karnezis, R. Shayan, S. B. Fox and M. G. Achen, *Nat. Rev. Cancer*, 2014, **14**, 159–72.

58 G. R. Lloyd, L. E. Orr, J. Christie-Brown, K. McCarthy, S. Rose, M. Thomas and N. Stone, *Analyst*, 2013, **138**, 3900–8.

59 K. S. Blum and R. Pabst, *J. Anat.*, 2006, **209**, 585–595.

60 R. V. Krstic, *Human Microscopic Anatomy: An Atlas for Students of Medicine and Biology*, Springer Science and Business Media, New York, 2013.

61 R. Miranda, J. D. Khoury and L. J. Medeiros, *Atlas of Lymph Node Pathology*, Springer S., 2013.

62 B. Bird, M. Romeo, N. Laver and M. Diem, *J. Biophotonics*, 2009, **2**, 37–46.

63 M. Wilson, E. L. Rosato, K. A. Chojnacki, I. Chervoneva, J. C. Kairys, H. E. Cohn, F. E. R. Sr and C. Berger, *J. Surg. Res.*, 2008, **146**, 11–15.

64 B. Kuo and D. Urma, *GI Motil. online*, 2006.

65 W. J. Loehr, Z. Mujahed, F. D. Zahn, G. F. Gray and B. Thorbjarnarson, *Ann. Surg.*, 1969, **170**, 232–238.

66 G. N. Hounsfield, *Br. J. Radiol.*, 1973, **46**, 1016–1022.

67 J. F. Schenck, *Med. Phys.*, 1996, **23**, 815–850.

68 L. K. Griffeth, *Proc. (Bayl. Univ. Med. Cent).*, 2005, **18**, 321–30.

69 M. H. Fazel Zarandi, M. Zarinbal and M. Izadi, *Appl. Soft Comput.*, 2011, **11**, 285–294.

70 N. S. Konstantinos, C. M. Nutting and C. Roussos, *Tumors of the Chest*, Springer Berlin Heidelberg, 2006.

71 K. M. Kerr, D. Lamb, C. G. Wathen, W. S. Walker and N. J. Douglas, *Thorax*, 1992, **47**, 337–341.

72 S. Kaminaka, H. Yamazaki, T. Ito, E. Kohda and H. Hamaguchi, *J. Raman Spectrosc.*, 2001, **32**, 139–141.

73 S. Campos, P. Davey, A. Hird, B. Pressnail, J. Bilbao, R. I. Aviv, S. Symons, F. Pirouzmand, E. Sinclair, S. Culleton, E. Desa, P. Goh and E. Chow, *Curr. Oncol.*, 2009, **16**, 62–6.

74 H. Kobayashi, M. R. Longmire, M. Ogawa, P. L. Choyke and S. Kawamoto, *Lancet Oncol.*, 2010, **11**, 589–95.

75 J. D. Horsnell, J. A. Smith, M. Sattlecker, A. Sammon, J. Christie-Brown, C. Kendall and N. Stone, *Surgeon*, 2012, **10**, 123–7.

76 C. Ozaslan and B. Kuru, *Am. J. Surg.*, 2004, **187**, 69–72.

77 M. H. Ross and W. Pawlina, *Histology A text and Atlas*, Lippincott Williams & Wilkins, 6th edn., 2011.

78 M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener, *IEEE Rev. Biomed. Eng.*, 2009, **2**, 147–171.

79 H. L. Ioachim and H. Ratech, *Lymph node pathology*, Lippincott Williams & Wilkins, PA, USA, 3rd ed., 2002.

80 R. A. Weinberg, *The biology of Cancer*, Garland Science, Taylor and Francis Group, LLC., 2007.

81 R. Küppers and M.-L. Hansmann, *Int. J. Biochem. Cell Biol.*, 2005, **37**, 511–7.

82 B. H. Tindle, *Calif. Med.*, 1972, **116**, 51–52.

83 C. Krafft, S. B. Sobottka, G. Schackert and R. Salzer, *Analyst*, 2004, **129**, 921–5.

84 S. S. Raab, D. M. Grzybicki, J. E. Janosky, R. J. Zarbo, F. a Meier, C. Jensen and S. J. Geyer, *Cancer*, 2005, **104**, 2205–13.

85 N. Stone, C. Kendall, J. Smith, P. Crow and H. Barr, *Faraday Discuss.*, 2004, **126**, 141.

86 S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander and T. R. Golub, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 15149–54.

87 C. Krafft, L. Shapoval, S. B. Sobottka, K. D. Geiger, G. Schackert and R. Salzer, *Biochim. Biophys. Acta*, 2006, **1758**, 883–91.

88 J. D. Kronz, W. H. Westra and J. I. Epstein, *Cancer*, 1999, **86**, 2426–35.

89 J. J. Turner, A. M. Hughes, A. Kricker, S. Milliken, A. Grulich, J. Kaldor and B. Armstrong, *Cancer Epidemiol. Biomarkers Prev.*, 2005, **14**, 2213–9.

90 N. L. Harris, E. S. Jaffe, J. Diebold, G. Flandrin and J. Vardiman, *Ann. Oncol.*, 2000, **11**, S3–S10.

91 M. Winter, R. Gibson, A. Ruszkiewicz, S. K. Thompson and B. Thierry, *Int. J. cancer*, 2014, **0**, 1–9.

92 M. M. Boddington and A. I. Spriggs, *Br. Med. J.*, 1965, 1523–1529.

93 S. Satyanarayana, A. Kalghatgi, A. Muralidhar, R. Prasad, K. Jawed and

A. Trehan, *Med. J. Armed Forces India*, 2002, **58**, 33–37.

94 F. Mayall, A. Darlington and B. Harrison, *J. Clin. Pathol.*, 2003, **56**, 821–825.

95 Q. He, X. Fan, T. Yuan, L. Kong, X. Du, D. Zhuang and Z. Fan, *Breast*, 2007, **16**, 303–6.

96 C. C. Cheung, B. Carydis, S. Ezzat, Y. C. Bedard and S. L. Asa, *J. Clin. Endocrinol. Metab.*, 2001, **86**, 2187–90.

97 M. W. H. Leenders, M. Broeders, C. Croese, M. C. Richir, H. L. S. Go, B. L. A. M. Langenhorst, S. Meijer and W. H. Schreurs, *The Breast*, 2012, **21**, 578–83.

98 S. E. Cross, Y.-S. Jin, J. Rao and J. K. Gimzewski, *Nat. Nanotechnol.*, 2007, **2**, 780–783.

99 L. M. Staudt, *N. Engl. J. Med.*, 2003, **348**, 1777–85.

100 D. Gonzalez de Castro, P. a Clarke, B. Al-Lazikani and P. Workman, *Clin. Pharmacol. Ther.*, 2013, **93**, 252–9.

101 T. Nielsen, F. Hsu and K. Jensen, *Clin. Cancer Res.*, 2004, **10**, 5367–5374.

102 R. Chang, *Basic Principles of Spectroscopy*, McGraw-Hill, Inc., 1971.

103 P. Atkins and J. DePaula, *Physical Chemistry for the Life Sciences*, Oxford University Press, 2nd edn., 2011.

104 D. I. Bower and W. F. Maddens, *The Vibration Spectroscopy of Polymers*, Cambridge University Press, 1989.

105 C. N. Banwell and E. M. McCash, *Fundamentals of Molecular Spectroscopy 4th ed.*, McGraw-Hill International Limited., London, 1994.

106 J. M. Hollas, *Modern Spectroscopy*, John Wiley and Sons, Ltd., 1987.

107 B. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*, John Wiley and Sons, Ltd., 2004.

108 R. L. McCreery, *Raman Spectroscopy for Chemical Analysis*, John Wiley and Sons, Inc., 2000.

109 R. S. Das and Y. K. Agrawal, *Vib. Spectrosc.*, 2011, **57**, 163–176.

110 D. R. Browning, *Spectroscopy*, McGraw-Hill Publishing Company Limited., 1969.

111 C. V. Raman and K. S. Krishnan, *Nature*, 1928, **121**, 501–502.

112 D. A. Skoog, F. J. Holler and S. R. Crouch, *Principles of Instrumental Analysis. 6th ed.*, Thompson Brooks/Cole., 2007.

113 Q. Tu and C. Chang, *Nanomedicine*, 2012, **8**, 545–558.

114 M. Wagner, *Annu. Rev. Microbiol.*, 2009, **63**, 411–429.

115 D. Ball, *Spectroscopy*, 2001, **16**, 32–35.

116 S. B. Howell, *Handbook of CCD astronomy*, Cambridge University Press, New York, 2nd ed., 2006.

117 J. C. de Mello, H. F. Wittmann and R. H. Friend, *Adv. Mater.*, 1997, **9**, 230–232.

118 Andor Technology, 2003.

119 C. Kallaway, L. M. Almond, H. Barr, J. Wood, J. Hutchings, C. Kendall and N. Stone, *Photodiagnosis Photodyn. Ther.*, 2013, **10**, 207–19.

120 I. R. Lewis and H. Edwards, *Handbook of Raman Spectroscopy: From Research Laboratory to the Process Line*, Marcel Dekker, Inc., 2001.

121 P. Vandenabeele, *Practical Raman Spectroscopy: An Introduction*, John Wiley & Sons, Ltd., Chichester, 2013.

122 Z. Huang, H. Zeng, I. Hamzavi, D. I. McLean and H. Lui, *Opt. Lett.*, 2001, **26**, 1782–1784.

123 T. Bocklitz, A. Walter, K. Hartmann, P. Rösch and J. Popp, *Anal. Chim. Acta*, 2011, **704**, 47–56.

124 L. M. Fullwood, D. Griffiths, K. Ashton, T. Dawson, R. W. Lea, C. Davis, F. Bonnier, H. J. Byrne and M. J. Baker, *Analyst*, 2014, **139**, 446–454.

125 E. B. Hanlon, R. Manoharan, T. W. Koo, K. E. Shafer, J. T. Motz, M. Fitzmaurice, J. R. Kramer, I. Itzkan, R. R. Dasari and M. S. Feld, *Phys. Med. Biol.*, 2000, **45**, R1–R59.

126 G. Shetty, C. Kendall, N. Shepherd, N. Stone and H. Barr, *Br. J. Cancer*, 2006, **94**, 1460–1464.

127 J. S. Thakur, H. Dai, G. K. Serhatkulu, R. Naik, V. M. Naik, A. Cao, A. Pandya, G. W. Auner, R. Rabah, M. D. Klein and C. Freeman, *J. Raman …*, 2007, **38**, 127–134.

128 N. O. B. Luttschwager, *Raman spectroscopy of conformational rearrangements at low temperatures: folding and stretching af alkanes in supersonic jets*, Switzerland, 2014.

129 J. Sawatzki, R. Fishcer, H. Scheer and F. Siebert, *Proc. Natl. Acad. Sci. U. S. A.*, 1990, **87**, 5903–5906.

130 R. a. Bitar Carter, A. a. Martin, M. M. Netto and F. a. Soares, *Biomed. Vib. Spectrosc.*, 2004, **5321**, 190–197.

131 F. Bonnier, S. M. Ali, P. Knief, H. Lambkin, K. Flynn, V. McDonagh, C.

Healy, T. C. Lee, F. M. Lyng and H. J. Byrne, *Vib. Spectrosc.*, 2012, **61**, 124–132.

132 F. Bonnier, A. Mehmood, P. Knief, a. D. Meade, W. Hornebeck, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. C. Lee, F. M. Lyng and H. J. Byrne, *J. Raman Spectrosc.*, 2011, **42**, 888–896.

133 C. Krafft, S. Dochow, I. Latka, B. Dietzek and J. Popp, *Biomed. Spectrosc. Imaging*, 2012, **1**, 39–55.

134 T. D. Wang and J. Van Dam, *Clin. Gastroenterol. Hepatol.*, 2004, **2**, 744–753.

135 G. J. Puppels, T. C. Bakker Schut, N. M. Sijtsema, M. Grond, F. Maraboeuf, C. G. de Grauw, C. G. Figdor and J. Greve, *J. Mol. Struct.*, 1995, **347**, 477–483.

136 R. Ziegler, *J. Nutr.*, 1989, 116–122.

137 F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

138 N. Huang, M. Short, J. Zhao, H. Wang, H. Lui, M. Korbelik and H. Zeng, *Opt. Express*, 2011, **19**, 22892.

139 I. Notingher, *Sensors*, 2007, **7**, 1343–1358.

140 R. E. Kast, G. K. Serhatkulu, A. Cao, A. K. Pandya, H. Dai, J. S. Thakur, V. M. Naik, R. Naik, M. D. Klein, G. W. Auner and R. Rabah, *Biopolymers*, 2007, **89**, 235–241.

141 K. Kneipp, H. Kneipp, I. Itzkan, R. R. Dasari and M. S. Feld, *Chem. Rev.*, 1999, **99**, 2957–2976.

142 J. T. Motz, M. Hunter, L. H. Galindo, J. a Gardecki, J. R. Kramer, R. R. Dasari and M. S. Feld, *Appl. Opt.*, 2004, **43**, 542–554.

143 ICNIRP, *Health Phys.*, 2013, **105**, 74–96.

144 M. G. Shim and B. C. Wilson, *J. Raman Spectrosc.*, 1997, **28**, 131–142.

145 M. G. Shim, B. C. Wilson, E. Marple and M. Wach, *Appl. Spectrosc.*, 1999, **53**, 619–627.

146 J. C. C. Day and N. Stone, *Appl. Spectrosc.*, 2013, **67**, 349–354.

147 I. E. Iping Petterson, J. C. C. Day, L. M. Fullwood, B. Gardner and N. Stone, *Anal. Bioanal. Chem.*, 2015, **407**, 8311–8320.

148 L. M. Almond, J. Hutchings, N. Shepherd, H. Barr, N. Stone and C. Kendall, *J. Biophotonics*, 2011, **4**, 685–95.

149 P. Matousek and N. Stone, *Analyst*, 2009, **134**, 1058–66.

150 P. Matousek, I. P. Clark, E. R. C. Draper, M. D. Morris, A. E. Goodship, N. Everall, M. Towrie, W. F. Finney and A. W. Parker, *Appl. Spectrosc.*, 2005, **59**, 393–400.

151 P. Matousek, E. R. C. Draper, A. E. Goodship, I. P. Clark, K. L. Ronayne and A. W. Parker, *Cent. Laser Facil. Annu. Rep.*, 2006, 133–135.

152 A. Barhoumi, D. Zhang, F. Tam and N. J. Halas, *J. Am. Chem. Soc.*, 2008, **130**, 5523–5529.

153 H. J. Byrne, G. Sockalingum and N. Stone, in *Biomedical Applications of Synchrotron Infrared Microspectroscopy*, 2011.

154 C. Matthäus, T. Chernenko, J. a Newmark, C. M. Warner and M. Diem, *Biophys. J.*, 2007, **93**, 668–73.

155 J. W. Chan, D. S. Taylor, T. Zwerdling, S. M. Lane, K. Ihara and T. Huser, *Biophys. J.*, 2006, **90**, 648–656.

156 B. C. Smith, *Fundamentals of Fourier Transform Infrared Spectroscopy*, CRC Press, Tayor and Francis Group, LLC, 2nd edn., 2011.

157 Thermo Nicolet, *Introduction to Fourier Transform Infrared Spectrometry*, Thermo Nicolet Corporation, 2001.

158 P. R. Griffiths and J. A. de Haseth, *Fourier Transform Infrared Spectrometry*, John Wiley & Sons, Inc., New Jersey, 2nd edn., 2007.

159 E. N. Lewis, P. J. Treado, R. C. Reeder, G. M. Story, a E. Dowrey, C. Marcott and I. W. Levin, *Anal. Chem.*, 1995, **67**, 3377–81.

160 N. Jaggi and D. R. Vij, in *Handbook of Applied Solid State Spectroscopy*, ed. V. D. R., Springer, New York, 2006.

161 K. M. Dorling and M. J. Baker, *Trends Biotechnol.*, 2013, **31**, 437–438.

162 G. J. Thomas Jr, *Annu. Rev. Biophys. Biomol. Struct.*, 1999, **28**, 1–27.

163 N. K. Afseth, V. H. Segtnan and J. P. Wold, *Appl. Spectrosc.*, 2006, **60**, 1358–67.

164 T. W. Randolph, *Cancer Biomark*, 2006, **2**, 135–144.

165 J. Popp, V. V. Tuchin, A. Chiou and S. H. Heinemann, *Handbook of Biophotonics, Volume 2: Photonics for Health Care*, Wiley-VCH Verlag & Co., Weinheim, Germany, 2012.

166 A. G. Ryder, G. M. O'Connor and T. J. Glynn, *J. Forensic Sci.*, 1999, **44**, 1013–1019.

167 H. Zha, X. He, C. Ding and H. Simon, *MIT Press*, 2001, 1057--1064.

168 M. J. Baker, E. Gazi, M. D. Brown, J. H. Shanks, P. Gardner and N. W. Clarke, *Br. J. Cancer*, 2008, **99**, 1859–66.

169 I. T. Jolliffe, *Principal component analysis*, Springer-Verlag New York, Inc., 2nd edn., 2002.

170 R. G. Brereton, *Chemometrics for Pattern Recognition*, John Wiley & Sons, 2009.

171 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.

172 G. R. Lloyd and N. Stone, *Appl. Spectrosc.*, 2015, **69**, 1066–73.

173 A. K. Akobeng, *Acta Paediatr.*, 2007, **96**, 338–341.

174 J. R. Hands, P. Abel, K. Ashton, T. Dawson, C. Davis, R. W. Lea, A. J. S. McIntosh and M. J. Baker, *Anal. Bioanal. Chem.*, 2013, **405**, 7347–7355.

175 K. Gajjar, L. D. Heppenstall, W. Pang, K. M. Ashton, J. Trevisan, I. I. Patel, V. Llabjani, H. F. Stringfellow, P. L. Martin-Hirsch, T. Dawson and F. L. Martin, *Anal. Methods*, 2013, **5**, 89–102.

176 C. Krafft, B. Belay, N. Bergner, B. F. M. Romeike, R. Reichart, R. Kalff and J. Popp, *Analyst*, 2012, **137**, 5533–7.

177 Y. Zhou, C.-H. Liu, Y. Sun, Y. Pu, S. Boydston-White, Y. Liu and R. R. Alfano, *J. Biomed. Opt.*, 2012, **17**, 116021.

178 L. M. Fullwood, G. Clemens, D. Griffiths, K. Ashton, T. P. Dawson, R. W. Lea, C. Davis, F. Bonnier, H. J. Byrne and M. J. Baker, *Anal. Methods*, 2014.

179 S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. . G. Yeoh and Z. Huang, *Br. J. Surg.*, 2010, **97**, 550–557.

180 M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Yan So and Z. Huang, *Int. J. cancer*, 2011, **128**, 2673–80.

181 M. G. Shim, L. Wong Kee Song, N. E. Marcon and B. C. Wilson, *Photochem. Photobiol.*, 2000, **72**, 146–150.

182 C. a Lieber and M. H. Kabeer, *J. Pediatr. Surg.*, 2010, **45**, 549–54.

183 A. Nijssen, T. C. Bakker Schut, F. Heule, P. J. Caspers, D. P. Hayes, M. H. A. Neumann and G. J. Puppels, *J. Invest. Dermatol.*, 2002, **119**, 64–69.

184 S. Sigurdsson, P. A. Philipsen, L. K. Hansen, J. Larsen, M. Gniadecka and H. C. Wulf, *IEEE Trans. Biomed. Eng.*, 2004, **51**, 1784–93.

185 H. Lui, J. Zhao, D. McLean and H. Zeng, *Cancer Res.*, 2012, **72**, 2491–2500.

186 M. Gniadecka, P. A. Philipsen, S. Sigurdsson, S. Wessel, O. F. Nielsen, D. H. Christensen, J. Hercogova, K. Rossen, H. K. Thomsen, R. Gniadecki, L. K. Hansen and H. C. Wulf, *J. Invest. Dermatol.*, 2004, **122**, 443–9.

187 M. Gniadecka, N. N. Mortensen, O. F. Nielsen and D. H. Christensen, *J. Ramna Spectrosc.*, 1997, **28**, 125–129.

188 C. a Lieber, S. K. Majumder, D. L. Ellis, D. D. Billheimer and A. Mahadevan-Jansen, *Lasers Surg. Med.*, 2008, **40**, 461–7.

189 K. Lin, D. L. P. Cheng and Z. Huang, *Biosens. Bioelectron.*, 2012, **35**, 213–7.

190 N. Stone, P. Stavroulaki, C. Kendall, M. Birchall and H. Barr, *Laryngoscope*, 2000, **110**, 1756–63.

191 M. Moreno, L. Raniero, E. Â. L. Arisawa, A. M. Espírito Santo, E. A. P. Santos, R. A. Bitar and A. A. Martin, *Theor. Chem. Acc.*, 2009, **125**, 329–334.

192 H. Abramczyk, J. Surmacki, B. Brożek-Płuska, Z. Morawiec and M. Tazbir, *J. Mol. Struct.*, 2009, **924–926**, 175–182.

193 H. Abramczyk, B. Brozek-Pluska, J. Surmacki, J. Jablonska and R. Kordek, *J. Mol. Liq.*, 2011, **164**, 123–131.

194 K. E. Shafer-Peltier, A. S. Haka, M. Fitzmaurice, J. Crowe, J. Myles, R. R. Dasari and M. S. Feld, *J. Raman Spectrosc.*, 2002, **33**, 552–563.

195 E. M. Kanter, E. Vargis, S. Majumder, M. D. Keller, E. Woeste, G. G. Rao and A. Mahadevan-Jansen, *J. Biophotonics*, 2009, **2**, 81–90.

196 P. Crow, a Molckovsky, N. Stone, J. Uff, B. Wilson and L.-M. WongKeeSong, *Urology*, 2005, **65**, 1126–30.

197 M. Tollefson, J. Magera, T. Sebo, J. Cohen, A. Drauch, J. Maier and I. Frank, *Br. J. Urol. Int.*, 2010, **106**, 484–488.

198 D. P. Lau, Z. Huang, H. Lui, C. S. Man, K. Berean, M. D. Morrison and H. Zeng, *Lasers Surg. Med.*, 2003, **32**, 210–4.

199 Z. Huang and A. McWilliams, *Int. J. cancer*, 2003, **107**, 1047–52.

200 S. Elmore, *Toxicol. Pathol.*, 2006, **34**, 425–54.

201 J. Smith, C. Kendall, a Sammon, J. Christie-Brown and N. Stone, *Technol. Cancer Res. Treat.*, 2003, **2**, 327–32.

202 J. Smith, C. Kendall, A. Sammon, J. Christie-Brown, T. Mandalia and N. Stone, in *Diagnostic Optical Spectroscopy in Biomedicine III*, 2005, vol. 5862.

203 J. Horsnell, P. Stonelake, J. Christie-Brown, G. Shetty, J. Hutchings, C. Kendall and N. Stone, *Analyst*, 2010, **135**, 3042–3047.

204 M. Sattlecker, C. Bessant, J. Smith and N. Stone, *Analyst*, 2010, **135**, 895–901.

205 L. E. Orr, J. Christie-Brown, J. C. Hutchings, K. McCarthy, S. Rose, M. Thomas and N. Stone, *Proc. SPIE*, 2010, **7548**.

206 B. Bird, K. Bedrossian, N. Laver, M. Miljković, M. J. Romeo and M. Diem, *Analyst*, 2009, **134**, 1067–1076.

207 Y. Liu, Y. Xu, Y. Zhang, D. Wang, D. Xiu, Z. Xu, X. Zhou, J. Wu and X. Ling, *Br. J. Surg.*, 2011, **98**, 380–4.

208 M. Isabelle, N. Stone and H. Barr, *Spectroscopy*, 2008, **22**, 97–104.

209 P. Andrus and R. Strickland, *Biospectroscopy*, 1998, **4**, 37–46.

210 P. Crow, J. S. Uff, J. A. Farmer, M. P. Wright and N. Stone, *BJU Int.*, 2004, **93**, 1232–1236.

211 F. S. Parker, *Applications of Infrared, Raman and Resonance Raman spectroscopy in Biochemistry*, Plenum Press, New York., New York, 1983.

212 Z. Movasaghi, S. Rehman and I. U. Rehman, *Appl. Spectrosc. Rev.*, 2007, **42**, 493–541.

213 B. Hernández, F. Pfluger, E. López-tobar, S. G. Krukliz, J. V Garcia-Ramos, S. Sanchez-Cortes and M. Ghomi, *Raman Spectrosc.*, 2014, **45**, 657–664.

214 I. Notingher, C. Green, C. Dyer, E. Perkins, N. Hopkins, C. Lindsay and L. L. Hench, *J. R. Soc. Interface*, 2004, **1**, 79–90.

215 A. C. S. Talari, Z. Movasaghi, S. Rehman and I. U. Rehman, *Appl. Spectrosc. Rev.*, 2014, **50**, 46–111.

216 L. E. Kamemoto, A. K. Misra, S. K. Sharma, M. T. Goodman, L. U. K. Hugh, A. C. Dykes and T. Acosta, *Appl. Spectrosc.*, 2010, **64**, 255–261.

217 R. Malini, K. Venkatakrishna, J. Kurien, K. M. Pai, L. Rao, V. B. Kartha and C. M. Krishna, *Biopolymers*, 2006, **81**, 179–193.

218 O. Tarhan, E. Tarhan and S. Harsa, *J. Dairy Res.*, 2014, **81**, 98–106.

219 R. K. Dukor, *Handb. Vib. Spectrosc.*, 2002, 3335–3361.

220 A. E. Aliaga, I. Osorio-Roman, P. Leyton, C. Garrido, J. Carcamo, C. Caniulef, F. Celis, G. Diaz, E. Clavijo, G.-J. J. S and M. M. Campos-vallette, *J. Raman Spectrosc.*, 2008, **40**, 164–169.

221 I. R. Martins Ramos, A. Malkin and F. M. Lyng, *Biomed Res. Int.*, 2015.

222 B. Hernández, F. Pfluger, S. G. Kruglik and M. Ghomi, *J. Raman*

Spectrosc., 2013, **44**, 827–833.

223 W. T. Cheng, M. T. Liu, H. N. Liu and S. Y. Lin, *Microsc. Res. Tech.*, 2005, **68**, 75–79.

224 J. T. Pelton and L. R. McLean, *Anal. Biochem.*, 2000, **277**, 167–176.

225 C. Krafft, L. Neudert, T. Simat and R. Salzer, *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, 2005, **61**, 1529–35.

226 R. J. Lakshmi, V. B. Kartha and C. M. Krishna, *Radiat. Res.*, 2002, **157**, 175–182.

227 S. L. Jacques, *Phys. Med. Biol.*, 2013, **58**, 5007–5008.

228 M. Z. Vardaki, B. Gardner, N. Stone and P. Matousek, *Analyst*, 2015.

229 a. J. Ruiz-Chica, M. a. Medina, F. Sánchez-Jiménez and F. J. Ramírez, *J. Raman Spectrosc.*, 2004, **35**, 93–100.

230 M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. a Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner and F. L. Martin, *Nat. Protoc.*, 2014, **9**, 1771–91.

231 C. Pond and C. Mattacks, *J. Lipid Res.*, 1995, **36**, 2219–2231.

232 P.-Y. von der Weid and K. J. Rainey, *Aliment. Pharmacol. Ther.*, 2010, **32**, 697–711.

233 Y. Wang and G. Oliver, *Genes Dev.*, 2010, **24**, 2115–26.

234 F. F. de Mul, H. Buiteveld, J. Lankester, J. Mud and J. Greve, *Hum. Pathol.*, 1984, **15**, 1062–8.

235 Y. S. Park, J. Lee, J. C. Pang, D. H. Chung, S.-M. Lee, J.-J. Yim, S.-C. Yang, C.-G. Yoo, Y. W. Kim and S. K. Han, *J. Korean Med. Sci.*, 2013, **28**, 550–4.

236 G. Murakami, *Ann. Surg. Oncol.*, 2004, **11**, 279S–284S.

237 M. Romeo and M. Diem, *Vib. Spectrosc.*, 2005, **38**, 115–119.

238 K. M. Nieman, I. L. Romero, B. Van Houten and E. Lengyel, *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*, 2013, **1831**, 1533–1541.

239 A. G. Renehan, M. Tyson, M. Egger, R. F. Heller and M. Zwahlen, *Lancet*, 2008, **371**, 569–78.

240 E. D. Rosen and B. M. Spiegelman, *Cell*, 2014, **156**, 1–46.

241 B. Bird, M. Miljković, N. Laver and M. Diem, *Technol. Cancer Res. Treat.*, 2011, **10**, 135–144.

242 C. Luparello, *J. Carcinog. Mutagen.*, 2013, **S13**.

243 R. A. Walker, *Breast cancer Res.*, 2001, **3**, 143–145.

244 A. Haka, *Proc. …*, 2005, **102**, 12371–6.

245 N. F. Boyd, G. a. Lockwood, J. W. Byng, D. L. Tritchler and M. J. Yaffe, *Cancer Epidemiol. Biomarkers Prev.*, 1998, **7**, 1133–1144.

246 A. V Rao and L. G. Rao, *Pharmacol. Res.*, 2007, **55**, 207–216.

247 T. C. Bakker Schut, G. J. Puppels, Y. M. Kraan, J. Greve, L. L. van der Maas and C. G. Figdor, *Int. J. Cancer*, 1997, **74**, 20–5.

248 T. R. Hata, T. A. Scholz, I. V Ermakov, R. W. Mcclane, F. Khachik, W. Gellermann and L. K. Pershing, *J. Invest. Dermatol.*, 2000, **115**, 441–448.

249 B. Sitter, T. F. Bathen and I. S. Gribbestad, in *Cancer Imaging: Instrumentation and Applications, Volume 2*, Academic Press, 2007, pp. 45–51.

250 C. G. Guzzo, W. N. Pachas, R. S. Pinals and M. L. Krant, *Cancer*, 1969, **24**, 382–387.

251 F. Gielen, J. Dequeker, A. Drochmans, J. Wildiers and M. Merlevede, *Br. J. Cancer*, 1976, **34**, 279–285.

252 P. long Li, X. Zhang, T. fu Li, L. li Wang, L. tao Du, Y. mei Yang, J. Li, H. yan Wang, Y. Zhang and C. xin Wang, *Clin. Chim. Acta*, 2015, **439**, 148–153.

253 C. M. Krishna, G. D. Sockalingum, R. A. Bhat, L. Venteo, P. Kushtagi, M. Pluot and M. Manfait, *Anal. Bioanal. Chem.*, 2007, **387**, 1649–1656.

254 R. Sathyavathi, A. Saha, J. S. Soares, N. Spegazzini, S. McGee, R. Rao Dasari, M. Fitzmaurice and I. Barman, *Sci. Rep.*, 2015, **5**, 9907.

255 M. Diem, A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljković, C. Krafft and J. Popp, *J. Biophotonics*, 2013, **6**, 855–886.

256 J. Nallala, M.-D. Diebold, C. Gobinet, O. Bouché, G. D. Sockalingum, O. Piot and M. Manfait, *Analyst*, 2014, **139**, 4005–4015.

257 C. Gobinet, D. Sebiskveradze, V. Vrabie, A. Tfayli, O. Piot and M. Manfait, 2008, 1–5.

258 D. C. Fernandez, R. Bhargava, S. M. Hewitt and I. W. Levin, *Nat. Biotechnol.*, 2005, **23**, 469–474.

259 K. H. Kim, F. Schwartz and E. Fuchs, *Proc. Natl. Acad. Sci. U. S. A.*, 1984, **81**, 4280–4.

260 J. M. Howard and C. Johnston, *Abdom. Imaging*, 2013, **38**, 233–243.

261 G. Marchal, R. Oyen, J. Verschakelen, J. Gelin, A. L. Baert and R. C. Stessenst, *J. Ultrasound Med.*, 1985, **4**, 417–419.

262 L. M. Weiss, *Lymph Nodes*, Cambridge University Press, 2008.

263 P. Meksiarun, B. B. Andriana, H. Matsuyoshi and H. Sato, *Sci. Rep.*, 2016, **6**, 37068.

264 P. Li, Y. Yin, D. Li, S. W. Kim and G. Wu, *Br. J. Nutr.*, 2017, **98**, 237–252.

265 P. C. Calder, *J. Nutr.*, 2006, **136**, 288–293.

266 G. Wu, *Amino Acids*, 2009, **37**, 1–17.

267 2nd National Report on Biochemical indicators of Diet and Nutrition in the American Population, *Fat-soluble vitamins and nutrients*, 2004.

268 J. F. Gillooly, A. Hein and R. Damiani, 2015, 1–28.

269 M. Fang, J. Yuan, C. Peng and Y. Li, *Tumor Biol.*, 2014, **35**, 2871–2882.

270 J. C. Sok, J. a Lee, S. Dasari, S. Joyce, S. C. Contrucci, a M. Egloff, B. K. Trevelline, R. Joshi, N. Kumari, J. R. Grandis and S. M. Thomas, *Br. J. Cancer*, 2013, **109**, 3049–56.

271 K. Czamara, K. Majzner, M. Z. Pacia, K. Kochan, A. Kaczor and M. Baranska, *J. Raman Spectrosc.*, 2015, **46**, 4–20.

272 Y. Manabe, S. Toda, K. Miyazaki and H. Sugihara, *J. Pathol.*, 2003, **201**, 221–228.

273 W. L. Lo, J. Y. Lai, S. E. Feinberg, K. Izumi, S. Y. Kao, C. S. Chang, A. Lin and H. K. Chiang, *J. Raman Spectrosc.*, 2011, **42**, 174–178.

274 H. Nishino, H. Tokuda, M. Murakoshi, Y. Satomi, M. Masuda, M. Onozuka, S. Yamaguchi, J. Takayasu, J. Tsuruta, M. Okuda, F. Khachik, T. Narisawa, N. Takasuka and M. Yano, *Biofactors*, 2000, **13**, 89–94.

275 R. Edge, D. J. McGarvey and T. G. Truscott, *J. Photochem. Photobiol. B Biol.*, 1997, **41**, 189–200.

276 T. T. Nguyen, C. Gobinet, J. Feru, S. Brassart-Pasco, M. Manfait and O. Piot, *Adv. Biomed. Spectrosc.*, 2013, **7**, 105–110.

277 M. Hilska, Y. Collan, J. Peltonen, R. Gullichsen, H. Paajanen and M. Laato, *Eur. J. Surg.*, 1998, **164**, 457–64.

278 D. Gao and S. Li, *Cancer Microenviron.*, 2013, **6**, 213–230.

279 T. Knösel, K. Schlüns, U. Stein, H. Schwabe, P. M. Schlag, M. Dietel and I. Petersen, *Neoplasia*, 2004, **6**, 23–28.

280 P. Di Ninni, F. Martelli and G. Zaccanti, *Opt. Express*, 2010, **18**, 26854–26865.

281 M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. So and Z. Huang, *Technol. Cancer Res. Treat.*, 2011, **10**, 103–12.

282 M. Nyberg, K. Ramser and O. a Lindahl, *Analyst*, 2013, **138**, 4029–34.

283 M. Nyberg, V. Jalkanen, K. Ramser, B. Ljungberg, a. Bergh and O. a. Lindahl, *J. Med. Eng. Technol.*, 2015, **39**, 198–207.

284 J. Desroches, M. Jermyn, K. Mok, C. Lemieux-Leduc, J. Mercier, K. St-Arnaud, K. Urmey, M.-C. Guiot, E. Marple, K. Petrecca and F. Leblond, *Biomed. Opt. Express*, 2015, **6**, 2380.

285 Andor Technology (Oxford Instruments), .